THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Systems Biology of Protein Secretion in Human Cells

– Multi-omics Analysis and Modeling of the Protein Secretion Process in Human Cells and its Applications

RASOOL SAGHALEYNI

CHALMERS
UNIVERSITY OF TECHNOLOGY

Systems and Synthetic Biology
Department of Biology and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021

Systems Biology of Protein Secretion in Human Cells

– Multi-omics Analysis and Modeling of the Protein Secretion Process in Human Cells and its Applications

RASOOL SAGHALEYNI

ISBN 978-91-7905-517-2
Serial number (Löpsnummer): 4984

Division of Systems and Synthetic Biology
Department of Biology and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021

Cover illustration: Protein secretion modeling, Rasool Saghaleyni 2021

Printed by Chalmers digitaltryck
Gothenburg, Sweden 202

Systems Biology of Protein Secretion in Human Cells:
Multi-omics Analysis and Modeling of the Protein Secretion Process in Human Cells and its Applications.

Rasool Saghaleyni
Department of Biology and Biological Engineering
Chalmers University of Technology

ABSTRACT

Since the emergence of modern biotechnology, the production of recombinant pharmaceutical proteins has been an expanding field with high demand from industry. Pharmaceutical proteins have constituted the majority of top-selling drugs in the pharma industry during recent years. Many of these proteins require post-translational modifications and are therefore produced using mammalian cells such as Chinese Hamster Ovary cells. Despite frequent improvements in developing efficient cell factories for producing recombinant proteins, the natural complexity of the protein secretion process still poses serious challenges for the production of some proteins at the desired quantity and accepted quality. These challenges have been intensified by the growing demands of the pharma industry to produce novel products with greater structural complexity and increasing expectations from regulatory authorities in the form of new quality control criteria to guarantee product safety.

This thesis focuses on different aspects of the protein secretion process, including its engineering for cell factory development and analysis in diseases associated with its deregulation. A major part of this thesis involved the use of HEK293 cells as a human model cell-line for investigating the protein secretion process by generating different types of omics data and developing a computational model of the human protein secretion pathway. I compared the transcriptomic profile of cell lines producing erythropoietin (EPO; as a model secretory protein) at different rates to identify critical genes that potentially contributed to higher rates of protein secretion. Moreover, by performing a transcriptomic comparison of cells producing green fluorescent protein (GFP; as a model non-secretory protein) with EPO producers, I captured differences specifically related to secretory protein production. I sought to investigate further the factors contributing to increased recombinant protein production by analyzing additional omic layers such as proteomics and metabolomics in cells that exhibited different rates of EPO production. Moreover, I developed a toolbox (HumanSec) to extend the reference human genome-scale metabolic model (Human1) to encompass protein-specific reactions for each secretory protein detected in our proteomics dataset. I could predict the top host cell proteins (HCPs) that compete with EPO for metabolic and energetic resources by generating cell-line specific protein secretion models and constraining the models using metabolomics data. Finally, based on the detected patterns of changes in our multi-omics investigations combined with a protein secretion sensitivity analysis using the metabolic model, I identified a list of genes and pathways that potentially play a crucial role in recombinant protein production and could serve as promising candidates for the targeted cell factory design.

In another part of the thesis, I studied the link between the expression profiles of genes involved in the protein secretory pathway (PSP) and various hallmarks of cancer. By

implementing a dual approach involving differential expression analysis and eight different machine learning algorithms, I investigated the expression changes in secretory pathway components across different cancer types to identify PSP genes whose expression was associated with tumor characteristics. I demonstrated that a combined machine learning and differential expression approach have a complementary nature and could highlight key PSP components relevant to features of tumor pathophysiology that may constitute potential therapeutic targets.

Keywords: protein secretion, integrative omics analysis, genome-scale modeling, protein secretion modeling, cancer protein secretory pathway, erythropoietin, HEK293

List of publications

This thesis is based on the following publications and manuscript:

Paper I: Malm, Magdalena*, Rasool Saghaleyni*, Magnus Lundqvist, Marco Giudici, Veronique Chotteau, Ray Field, Paul G. Varley, et al. 2020. "Evolution from Adherent to Suspension: Systems Biology of HEK293 Cell Line Development." *Scientific Reports* 10 (1): 18996.
*Authors contributed equally to this work
PMCID: PMC7642379

Paper II: Rasool Saghaleyni, Magdalena Malm, Jan Zrimec, Ronia Razavi, Num Wistbacka, Veronique Chotteau, Diane Hatton, et al. 2020. "Transcriptome Analysis of EPO and GFP HEK293 Cell-Lines Reveal Shifts in Energy and ER Capacity Support Improved Erythropoietin Production in HEK293F Cells." bioRxiv. [manuscript; under review].

Paper III: Rasool Saghaleyni, Magdalena Malm, Jonathan L. Robinson, Jan Zrimec, Veronique Chotteau, Diane Hatton, et al. 2020 "Multi-omics analysis of protein secretion in HEK293 cells suggests mTORC1 activation modulates higher recombinant Erythropoietin production. [manuscript]

Paper IV: Rasool Saghaleyni, Azam Sheikh Muhammad, Pramod Bangalore, Jens Nielsen, and Jonathan L. Robinson. 2021. "Machine Learning-Based Investigation of the Cancer Protein Secretory Pathway." *PLoS Computational Biology* 17 (4): e1008898.
PMCID: PMC8049480

Contribution summary

Paper I, II & III. Contributing to the research design, analysis of generated data and finding the key points in results, writing the original draft paper, and contributing to the review and editing process.

Paper IV. Contributed to data curation, formal analysis, investigations, finding and discussing the results, and wrote parts of the manuscript.

Preface

This dissertation is submitted for the partial fulfillment of the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at the Chalmers University of Technology. It is based on the work carried out between August 2016 and August 2021 in the Systems and Synthetic Biology division under the supervision of Prof. Jens Nielsen. The research was funded by the Knut and Alice Wallenberg Foundation, AstraZeneca, Swedish Foundation for Strategic Research (SSF), Swedish innovation agency Vinnova through AAVNova, CellNova and AdBIOPRO, and the Novo Nordisk Foundation.

Rasool Saghaleyni
August 2021

Table of contents

Background 1
The protein secretion process 1
The protein secretory pathway in human cells 2
Translocation of proteins to ER 3
Co-translational translocation 3
Post-translational translocation 3
Tail-anchored translocation 4
Protein folding and post-translational modifications in the ER 4
Quality control mechanisms in the ER 5
Targeting matured proteins to the Golgi by COPII-vesicles 5
Post-translational modifications in the Golgi 5
Transport from the Golgi to the final destination 6
Differences in the protein secretion process between yeast and mammalian cells 6
Pharmaceutical proteins are among the top-selling drugs 7
Host cells for producing pharmaceutical proteins 8
Bacteria and yeast 9
CHO cells 9
HEK293 cells 10
Deficiency in the protein folding process leads to proteopathy diseases 11
Current approaches in systems biology of protein secretion 12
Omics techniques and computational data analysis approaches 13
Genomics 14
Transcriptomics 14
Mass spectrometry-based omics approaches: proteomics and metabolomics 15
Integrative omics analysis 16
Concatenation-based integration 17
Model-based integration 17
Transformation-based integration 18
Modeling approaches 18
Genome-scale metabolic models 18
Principles 19
Reconstruction 20
Human1 21
Further GEM improvements 21
Protein secretion modeling 22

Aim and significance 24

Results & discussion 25
Key genes in cell morphology transformation (Paper I) 25
Genomic differences between HEK293 parental and progeny cell lines 26
Transcriptomic comparison of HEK293 cells 28

Cell demands for the production of secretory and non-secretory proteins are different (Paper II) 31
Energy availability is necessary for recombinant protein production 32
Ribosomal components adapt to meet protein production requirements 33
Mitochondrial ribosomal genes positively correlate with EPO secretion 34
Multi-omics of protein secretion by HEK293 cells (Paper III) 37
Multi-omics analysis of cells producing EPO at different rates 37
Investigating variables in principal component analysis 39
Predicting host cell proteins that compete with EPO secretion 41
Protein Secretion modeling by HumanSec toolbox 41
Protein specific information matrix (PSIM) 42
Generating a list of template reactions for the secretory pathway 42
Reconstruction of the secretory pathway in human cells 42
Constraint-based analysis of reconstructed secretory models 43
Host cell proteins competing for most with EPO 43
Analysis of the protein secretory pathway in cancer (Paper IV) 45
Machine learning analysis of the cancer transcriptome 45
Machine learning approaches detect PSP genes that are regulated by P53 46
PSP genes associated with malignant transformation 49
PSP genes associated with tumor stages 50
Future Perspectives 55
Conclusion 57
References 59

Abbreviations

ATP	adenosine triphosphate
CHO	Chinese hamster ovarian
DE	differentially expressed
DEA	differential expression analysis
EC-GEM	enzyme-constrained genome-scale metabolic model
ER	endoplasmic reticulum
FBA	flux balance analysis
GEM	genome-scale metabolic model
ML	machine learning
mRNA	messenger RNA
PCA	principal component analysis
PSP	protein secretory pathway
PTM	post-translational modification
r-protein	recombinant protein
secGEM	secretory genome-scale metabolic model

Acknowledgments

Many people and institutions directly or indirectly supported my Ph.D. project. I will attempt to acknowledge here some of which. The Swedish state and Chalmers University enabled this project by providing an excellent and productive research environment, Tack så mycket, Sverige! Thank you, Jens, for providing the opportunity for me to continue my research on what passionates me. For your continued trust, for your support when projects face challenges, and for your patience during my learning lag phase. Jonathan, I learned many things from you, in both science and soft skills. Thank you for your support, your patience, and your generosity. I still have much to learn from your problem-solving manner of thinking and efficiency. Thomas, thank you for the constructive discussions that we had in the early years of my Ph.D. Jan, thank you, it was a big chance for me to meet you and learn from you. Thank you, Johan, Magdalena, Diane, Luigi, and the rest of the collaborators in joint projects with the department of protein science at KTH university and the cell line development and engineering department at AstraZeneca. I learned many things during our collaboration, and I believe the results of our projects could greatly improve the process of production of recombinant proteins in mammalian cells. Adil, Amir, Rasmus, Leif, Pouyan, and Cheng, thank you for all the fruitful meetings we had at the start of my Ph.D. to share the field's latest findings and suggest available tools and approaches for performing the analysis in my projects. Thank you to members of the HMA subgroup for providing helpful suggestions. Thank you, Yu, Hao, Qi, and many others mentioned in this section for proofreading the thesis. Thank you, Anne-Lisa, Martina, Erica, Gunilla, Josefine, and the rest of the Bio department administration team for making a calm and lovely environment with your smooth management. Thank you, Jens Christian, Avlant, Carl, Gang, and other friendly people who were my officemates in both the first three years and two last years of my Ph.D. Johan, thank you for patiently hearing my questions and searching in Swedish content when I was confused about some official processes in Sweden. Skatteverket and I appreciate you. Paulo, thank you for introducing Gothenburg and Sysbio to me in the first week of my stay in Sweden. Mihail, Pierre, Dimitra, Hao, Sinisa, and other residents of Sysbio first floor, thank you for lunchtimes, table tennis matches, and organizing skill workshops. I enjoyed all of them. Thank you, Dina, and all of the other core value team members, for structuring an environment that all of us voluntarily share, respect, and contribute to science. Yasaman and Fariba, it was very nice to be in the same group with you, we shared many scientific and cultural sides, and it was always good to hear how you think about different things. Thank you, Parizad, Shaq, Shadi, Naghmeh, and other Iranian members of Sysbio, for sharing our culture, costumes and occasionally enjoying the excellent sense of talking in our native language. I gratefully acknowledge the support that I received from my family and friends during my doctoral studies. Maman, Baba, thank you for your never-ending support and love. Sister and brothers, I am always proud of you and I highly seek our reunion. Thank you to all my friends in Iran and other parts of the world. Yasaman(s), Rojin, Mohsen, Ehsan, Amir, Ali, and Mohammad, you made it easy to plan for every weekend and celebration times. I am sure we will stick together in the future as we stayed together in all ups and downs in the past five years. Thank you Bijan, for joining and planning running/hiking sessions enriched by sharing exciting books and podcasts; it was a great joy after packed working days.

You are the endless source of love and care, Setareh! Thank you for being my star through it all;

I love you.

Background

In the study of living organisms, systems biology considers the interactions between components of a biological entity and interprets the whole system's behavior in light of these interactions (Tavassoly, Goldfarb, and Iyengar 2018). Thus, systems biology as a scientific field could have different beginnings depending on the context and the scope of system-level thinking. However, considering specifically the term "Systems Biology" shows its emergence during the mid-sixties (Rosen 1968). In that time, the development of novel molecular biology methods and the emergence of high-throughput molecular identification approaches helped in collecting data from various biological pathways and enabled scientists to more precisely study cellular phenomena (Schena et al. 1995; Gygi et al. 1999). Soon, many scientists considered the need for holistic analyzing approaches that could implement all available data to study complex biological pathways and understand the underlying biological mechanisms.

The protein secretion process is among the most complicated pathways within cells. In human cells, more than 500 proteins support the core functionality of the secretory pathway for the production and processing of proteins (Feizi et al. 2017). This complexity highlights the need to apply a systematic approach in studying protein secretion through a holistic discipline by considering all network components.

The protein secretion process

Depending on their role and function, proteins can exist in diverse subcellular locations (Yang et al. 2014). There are different mechanisms that cells use to localize newly synthesized proteins (Park et al. 2011). The classical pathway involves orientating proteins to their appropriate destination using a signal peptide sequence on the primary protein, which is then recognized by transport proteins (Russell and Keiler 2007). Based on the signal sequence, various transport proteins are specific for different organelles or locations within the cell (Gomez-Navarro and Miller 2016). Correct sorting of proteins is crucial for cell viability since the failures in this process may lead to cellular dysfunction and disease (Hung and Link 2011). For example, Zellweger syndrome, Adrenoleukodystrophy (ALD), Refsum disease, and Parkinson's disease are associated with a failure to localize one or more specific proteins properly; either within the cells of a single tissue or in the whole body (Brown and Breton 2000).

Although most proteins are localized in the intracellular space(The UniProt Consortium 2017), some proteins possess a signal peptide guiding them to pass through the cell membrane and be exported outside the cell. These proteins are called secretory proteins and can have a wide range of roles for interacting with their surrounding environments (Uhlén et al. 2019). For example, hormones and signal peptides facilitate communication with other cells, antimicrobial peptides defend against invading factors, and digestive enzymes break down environmental resources and prepare them for uptake. The roles of secretory proteins are thus essential for cell physiology. Furthermore, understanding the protein secretion pathway is necessary for synthesizing and translocating desired secretory proteins with great economic/pharmaceutical value.

Out of the approximately 20,000 protein-coding genes in the human genome, more than one third (7,729) are predicted to either have a secretory signal peptide and be secreted out of the cell or have a transmembrane domain and be inserted into the cell membrane (Uhlén et al. 2015; Thul et al. 2017). The process of protein secretion begins with transferring newly synthesized secretory proteins into the endoplasmic reticulum (ER), followed by subsequent modification in the Golgi apparatus and exporting via secretory vesicles (Alberts 2017). Thus, there are many steps across different cellular compartments, from translation initiation to releasing secretory proteins outside the cell (**Figure 1**).

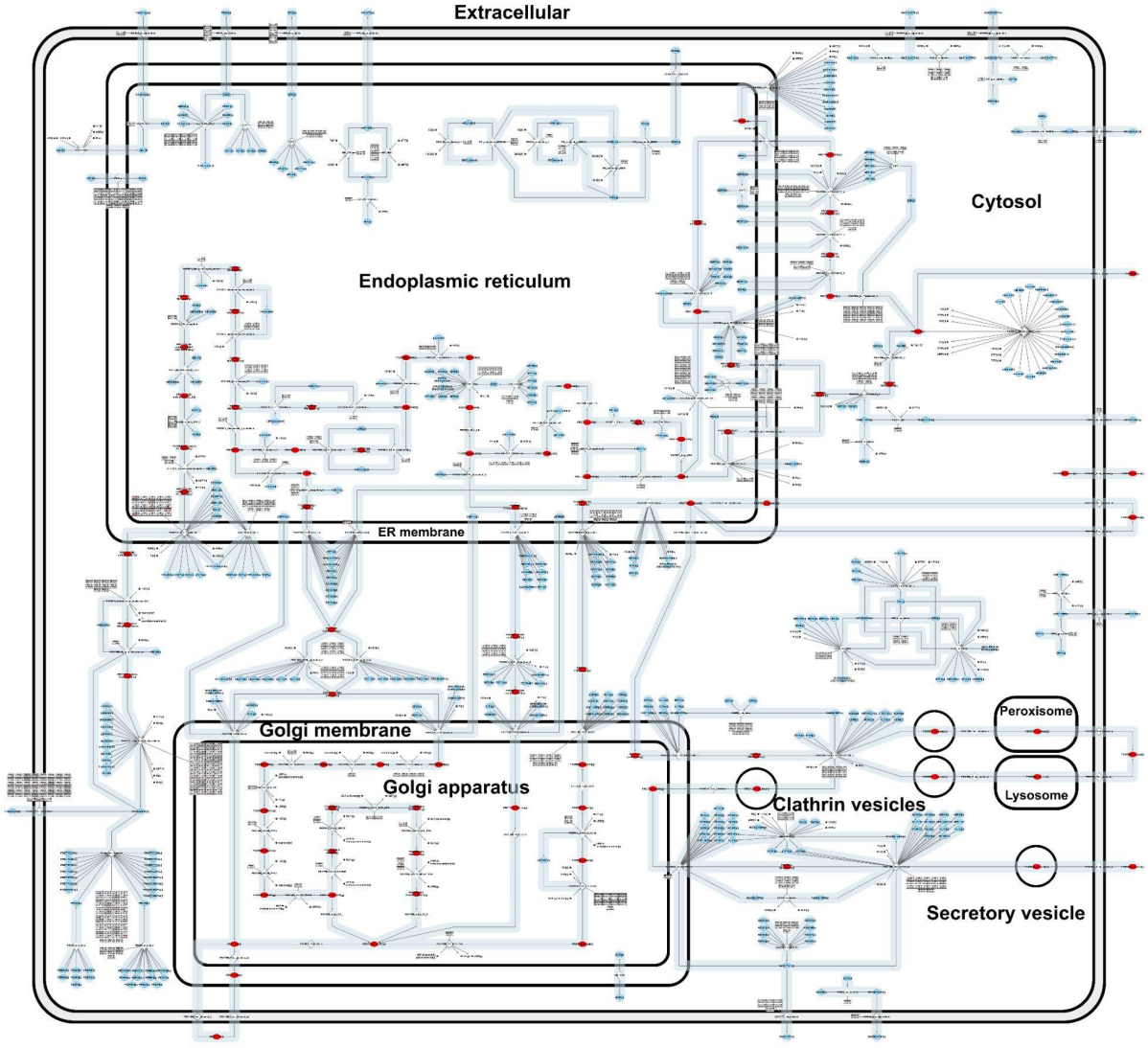

Figure 1. Metabolic map of the protein secretion process in human cells. The protein secretion process is a cross-compartmental process within the cell involving many components and tight connections with core metabolism. This map is available in high resolution at metabolicatlas.org.

The protein secretory pathway in human cells

Although secretory pathways are highly conserved from an evolutionary perspective (Mahlab and Linial 2014), there are important differences between yeast and mammalian cells (Feizi et al. 2013). Here I summarize the general secretory process in human cells, and in the next section, I will discuss differences in secretory pathways among bacteria, yeast, and human cells.

Translocation of proteins to ER

For most secretory proteins, the ER serves as a gateway to enter into the secretory process. As mentioned in the previous section, the main trigger for targeting a newly synthesized protein to the secretory pathway is the interaction between the signal peptide and its recognition particle (SRP). The signal peptide guides secretory proteins and targets them into the ER, where folding and post-translational modifications (PTMs) occur (Aviram and Schuldiner 2017). To transport a target protein across the ER membrane, there are three main mechanisms (Fewell and Brodsky 2013):

- Co-translational translocation
- Post-translational translocation
- Tail-anchored translocation

Co-translational translocation

Because the signal peptide in newly synthesized proteins is on the amino terminus of the protein, co-translational translocation starts as soon as the signal peptide emerges from the ribosome in the early translation steps (Lars Ellgaard et al. 2016). After SRP recognizes the signal peptide, the process of translation pauses, and the ribosome-protein complex is transferred to an SRP receptor on the ER. The process of signal peptide recognition by SRP is GTP-dependent. Binding SRP to SRP receptors on ER membranes causes ribosome-protein complexes to sit on translocon protein in the ER membrane (Skach 2007). Translocon serves as a conducting channel and is composed of the Sec61 translocation complex, one of the highly conserved proteins present in all domains of life (Gogala et al. 2014). This protein complex transports proteins into the ER in eukaryotes and out of the cell in prokaryotes (Gogala et al. 2014). As soon as the signal peptide of the nascent protein has been translocated into the translocon, signal peptidase in the ER membrane will cleave the signal peptide (S. J. Walker and Lively 2013). After signal peptide cleavage, the remaining protein continues entering into the ER lumen and the folding process begins to take place simultaneously. Chaperon proteins first cover the newly synthesized protein (Beissinger and Buchner 1998). Chaperons mainly inhibit the aggregation of newly synthesized proteins and prevent them from forming a nonfunctional structure (Beissinger and Buchner 1998). These proteins also assist with covalent folding and correct assembly of nascent protein, in addition to many other supportive functions (Saibil 2013).

Post-translational translocation

For a few secretory proteins and most of the proteins targeting other organelles like the nucleus and mitochondria, translocation occurs after translation. In post-translational translocation, the Sec61 protein in the ER membrane recognizes the signal peptide of the newly translated protein (Johnson, Powis, and High 2013). After signal peptide recognition by Sec61 and passing the protein through its lumen, the signal peptidase protein cleaves the signal peptide. Sec61 is located near to Sec62/Sec63 complex in the ER membrane, which acts as an activator for the ATPase activity of BiP proteins. BiP proteins bind non-specifically to proteins entering the ER lumen and prevent them from sliding back into the cytosol (Johnson, Powis, and High 2013).

Tail-anchored translocation

The last pathway for the translocation of proteins into the ER is tail-anchored translocation. This route is also SRP-dependent, but in this case, SRP recognizes the signal peptide located very close to the carboxy-terminal of the protein. Because the signal peptide is close to the carboxy-terminal, the ribosome must first release the protein before targeting the protein to the ER (Hegde and Keenan 2011).

Protein folding and post-translational modifications in the ER

Many proteins in the ER begin post-translational modification of nascent protein as soon as the protein enters into the ER (Wilk-Blaszczak n.d.). However, protein folding also occurs after entering the ER and after signal cleavage (Stevens and Argon 1999). Therefore, it is crucial to recognize that protein folding and post-translational modification do not occur in strict sequential order (Monte, del Monte, and Agnetti 2014). Protein modification can influence protein folding, which can, in turn, affect available sites for protein modification (Monte, del Monte, and Agnetti 2014). This phenomenon is very well explained and visualized in the review by (Braakman and Hebert 2013). The process and interaction between folding and modifying a protein will continue until it fully translocates into the ER (Mahlab and Linial 2014). Afterward, other post-translational modifications will continue to prepare for transportation into the Golgi apparatus (Alberts et al. 2002a), which often occurs through budding transport vesicles (Alberts et al. 2002a). Alternatively, there are some mechanisms in the ER for retro-translocation from ER to cytosol in case a protein fails to attain its desired conformation through the folding steps (Tsai, Ye, and Rapoport 2002). These proteins receive a ubiquitin tag and will be directed to proteasomes for degradation (Hochstrasser 1996).

Protein folding is highly dependent on spatial conditions. As soon as a protein has enough space, interactions between different protein domains will begin (McLeish 2005). On the other hand, having enough space in an environment with a high number of similar proteins or proteins with similar domains will dramatically increase the risk of misfolding and aggregation (McLeish 2005). Cells use heat shock proteins (HSPs) as chaperons and BiP to reduce this risk (Pobre, Poet, and Hendershot 2019). The other evolutionary solution to decrease the risk of protein misfolding is limiting the translation rate relative to folding and post-translational modification (Alexander et al. 2019). The translation rate for eukaryotic proteins is 4 to 5 residues per second. At the same time, in prokaryotic cells, it is much higher and is attributed as one of the reasons that eukaryotic proteins do not fold properly in prokaryotic hosts (Ross and Orlowski 1982).

Many secretory proteins carry N-linked glycans. The majority of modifications occur in the ER lumen by the specific machinery designed for this purpose (Aebi 2013). PTMs are usually added co-translationally following synthesis after 15 amino acids have entered into the ER lumen (Braakman and Hebert 2013). The glycosylation sites on the protein sequence are usually on an Asn residue in an Asn-X-Ser/Thr sequence (Lowenthal et al. 2016). Oligosaccharyltransferase is the enzyme responsible for transferring a glycan structure from ER membrane to an Asn residue in the nascent protein. Glycan structures are synthesized through the dolichol synthesis pathway in the ER membrane (Burda and Aebi 1999). The N-linked glycosylation is necessary for the

correct folding of proteins and increasing the stability of the protein (Hanson et al. 2009). From a conformational perspective, most glycans are on the protein's surface but can also be added to other locations on the protein structure (Lowenthal et al. 2016). Sites of glycosylation are highly protein-dependent (An, Froehlich, and Lebrilla 2009). Besides glycosylation sites, the total number of glycosylations on a protein surface could also be an essential feature for keeping a protein functional and stable (An, Froehlich, and Lebrilla 2009).

Quality control mechanisms in the ER

To exit from the ER, proteins should be properly folded and assembled, particularly if they are members of a multi-subunit complex (Araki and Nagata 2011). Other proteins without these qualifications still need to remain in the ER to be appropriately processed and folded (Liu et al. 2018). Persistently, misfolded proteins will be retrolocated to the cytosol for proteasomal degradation (Zhai et al. 2020). The quality control step is the main step to ensure that released proteins are correctly folded and functional. Unfortunately, most proteins that do not pass the quality criteria fail at this step (Alberts et al. 2002a). Because this step is essential for ensuring correctly folded and functional proteins, any failures in the quality check process can lead to different types of diseases (Yoshida 2007).

Targeting matured proteins to the Golgi by COPII-vesicles

After ER, proteins are transferred to the Golgi apparatus for further modifications such as O-linked-glycosylation and followed by final protein sorting (Alberts 2017). Vesicles responsible for this transfer from ER to Golgi are COPII vesicles (Jensen and Schekman 2011). COPII vesicles bud from specialized sites in the ER membrane and leave the ER to reach cis Golgi (closer site of Golgi to ER) (Verissimo and Pepperkok 2013). During this process, many quality control steps were assumed to guarantee that only correctly folded proteins leave the ER through this pathway (L. Ellgaard and Helenius 2001). However, new findings highlight that all the proteins, even ER-resident proteins, can leak out from ER to cis Golgi (Alberts et al. 2002a). Secretory proteins have a higher chance of transferring from ER to Golgi due to the presence of signal peptide for secretion ("Hematology" n.d.). After COPII vesicles leave the ER and reach the Golgi, they fuse to the cis Golgi to deliver their contents (Jensen and Schekman 2011). Soon after COPII vesicles fuse with the Golgi, COPI vesicles start budding from cis Golgi to return to the ER (Hsu, Lee, and Yang 2009). It should be noted that COPI vesicles transfer misfolded proteins back to the ER that has been brought to the Golgi by COPII vesicles (Hsu, Lee, and Yang 2009). The two vesicle types thus constitute a highly dynamic but controlled process orchestrated by signal transduction.

Post-translational modifications in the Golgi

The Golgi apparatus consists of compact stacks located close to the nucleus-far site of the ER (Glick 2000). It has been shown that the structural conformation of the nucleus, ER, Golgi, and vesicles that move between the ER and Golgi is organized by tubular connections that act as an infrastructure to maintain this conformation (Presley et al. 1997). Oligosaccharide chains are processed in the Golgi (Faye et al. 1986). The initial modifications for many N-linked oligosaccharides first happen in the ER and continue in Golgi (Helenius and Aebi 2001). Besides oligosaccharides, the other necessary modification in the Golgi apparatus is O-linked

glycosylation performed by a series of reactions catalyzed by glycosyltransferase enzymes (Stanley 2011). The complexity of pathways related to post-translational modifications, specifically different types of glycosylation, arises from their variable nature (Goto 2007). Whereas other macromolecules like proteins and nucleic acids are synthesized based on a template, large and complex glycosylation structures with up to 20 sugar units are produced based on the condition of the cell at the moment the protein passes through the ER and Golgi (Goto 2007). Currently, a big challenge in the pharmaceutical industry is overcoming the variability of these glycosylations between product batches. This complex nature of glycosylation structure formation is also a major challenge in modeling protein secretion processes (Wong 2005).

Transport from the Golgi to the final destination

Exocytosis is the process by which vesicles are delivered from the trans-Golgi to the cell membrane and secrete proteins into the extracellular environment (Söllner 2003). Depending on their signal, proteins will be sorted to their final destination after proper packaging in trans Golgi. Released vesicles from trans Golgi will direct to different cellular compartments or will be secreted to extracellular space (Guo, Sirkis, and Schekman 2014). After packaging secretory vesicles, motor proteins use their ATPase activity to push secretory vesicles along microtubules to reach the planned destination (Hunt and Stephens 2011). Transporting the proteins from trans Golgi to the plasma membrane is not the only function of secretory vesicles; recent studies have shown that, for some secretory proteins, a part of protein maturation occurs in the trans-Golgi and continues within secretory vesicles (Losev et al. 2006). Versus exocytosis, cells regulate the speed of endocytosis to allocate the required time for proper folding of the content of secretory vesicles (Alberts et al. 2002b).

Differences in the protein secretion process between yeast and mammalian cells

Yeast and mammalian cells have a similar structure and steps for the protein secretion process (Sakaguchi 1997). However, although the main procedures including co-translational translocation, folding, and N-glycosylation in ER, O-glycosylation in Golgi, ER-associated degradation, and final protein sorting are very similar between yeast and mammalian cells, there are still some differences that need to be considered. Here, I summarize the main differences that are known below:

1. Detection of proteins for ER-associated degradation is based on the pattern of protein post-translational modifications, which differs between yeast and mammalian cells (Eldeeb et al. 2019; Dunn 2003).
2. During translocation of proteins to the ER, the main chaperones acting in mammalian cells are BiP proteins, whereas the counterpart in the yeast cells is Kar2 (Kawaguchi and Ng 2011).
3. The quality control system in mammalian ER tags misfolded proteins with polysaccharide tags, while this mechanism in yeast cells is different (Xu and Ng 2015).

4. Components of COPII vesicles in mammalian cells contain four isoforms of Sec24, but in yeast cells, there are Sfb2/Sfb33 as homologs of Sec24 in mammalian cells.

Pharmaceutical proteins are among the top-selling drugs

In addition to the importance of the protein secretion pathway from a molecular biology perspective, broad applications of this pathway in industry and medicine further motivate its study. For example, in 2020, protein-based drugs constituted almost 25 percent of all new drugs approved by the FDA (the United States food and drug administration)* (Mullard 2021). Moreover, most drugs with the highest annual revenue are protein-based (8 out of 10 in 2018, Figure 2A) (Mullard 2021).

Protein drugs are mainly categorized as monoclonal antibodies, hormones, enzymes, blood factors, vaccines, antibiotics, and cytokines (Bruno, Miller, and Lim 2013). Peptides and proteins can target a broader range of molecules, making them powerful drug candidates for targeting different components in a biological network (Lau and Dunn 2018). Moreover, it is possible to consider more than one interaction between the protein drug and its target in the drug design process. More specific interactions increase the selectivity of the drug and prevent potential side effects by decreasing the off-target interactions with other unwanted molecules (Bruno, Miller, and Lim 2013).

Meanwhile, therapeutic proteins have a more complex structure and higher molecular weight than small molecule drugs (< 1 kDa for small molecules compared to 1 to > 10 kDa for larger proteins). The larger size can cause higher instability of these proteins than small molecule drugs (Bruno, Miller, and Lim 2013). Furthermore, more increased instability directly causes higher immunogenicity and a shorter half-life (Pisal, Kosloski, and Balu-Iyer 2010). Various strategies are under development to overcome these issues, such as adding chemical modifications to synthesized proteins and applying new formulation approaches. However, the production of pharmaceutical proteins in itself requires complex manufacturing setups, and adding more steps will further complicate the manufacturing process (Morrow and Felcone 2004).

---

*Comparing the list of new FDA-approved drugs from 2015 to 2020 with the same list from 2019 and 2020 indicates the observed increase in the number of protein-based drugs has not been because of granting emergency approvals to COVID19 (coronavirus disease 2019) proteins-based drugs.

The majority of the biologics currently on the market are produced through a cell culture fermentation process followed by downstream purification steps. The manufacturing process has a significant effect on the final quality of the product. Different parameters could be optimized to achieve higher quality and quantity in the protein production process, such as optimizing the bioreactor system, culturing media, purification, and removing impurities (Y. Ma, Lee, and Park 2020; Giuliani et al. 2011; Raynal et al. 2014). However, one component of this process significantly impacts the quality and quantity of the final product is the host cell line.

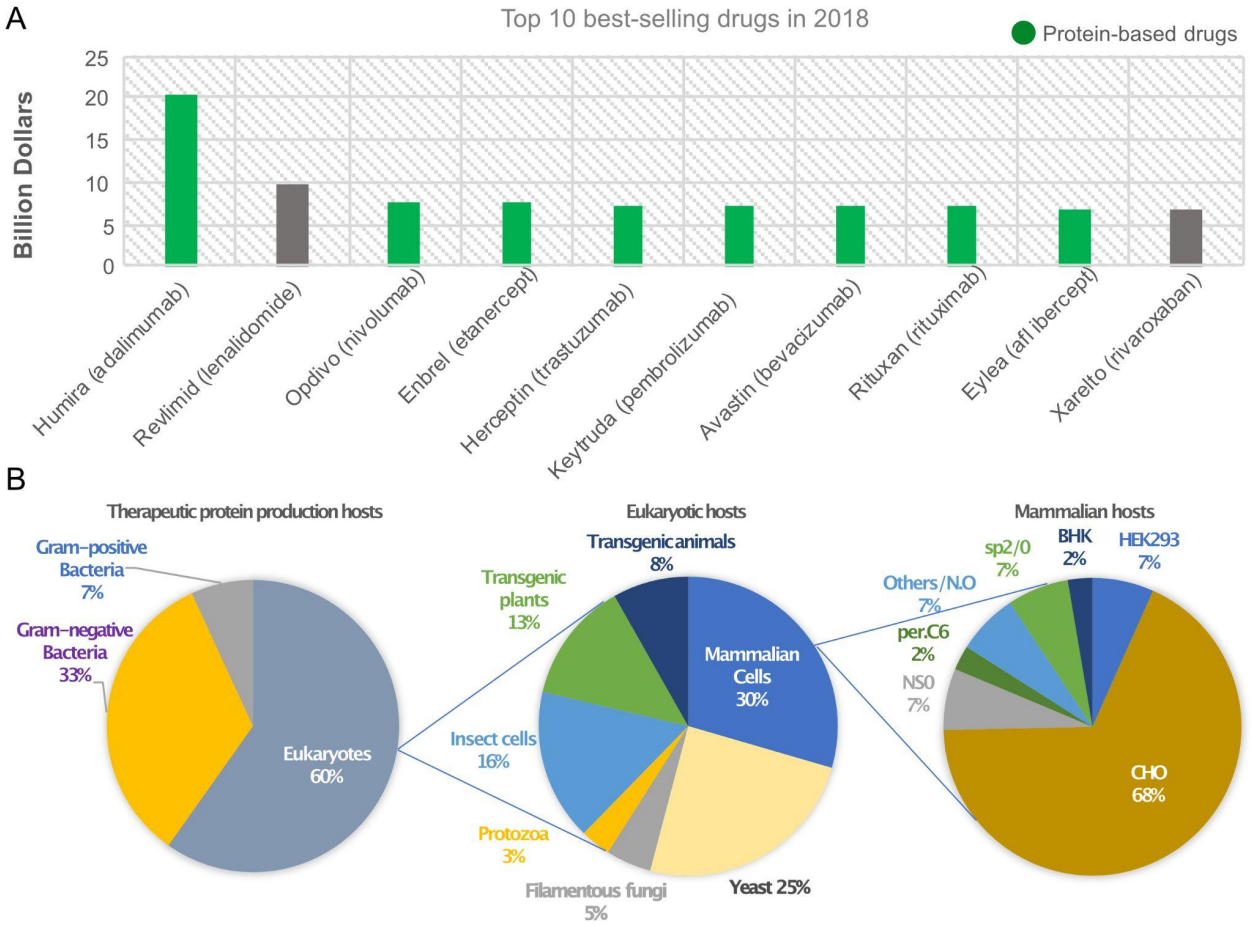

Figure 2. Pharmaceutical proteins are among the top-selling drugs. (A) Eight of the top 10 selling drugs in 2018 are protein-based. (B) Host cell lines for the production of pharmaceutical proteins. Eukaryotes produce more than 60% of pharmaceutical proteins, and currently, CHO cells are the preferred host for the production of more advanced products.

Host cells for producing pharmaceutical proteins

There are many different hosts available for producing recombinant proteins (Dumont et al. 2016). However, the choice of the most appropriate expression system for making one specific protein depends on its characteristics, the ability of the host cell-line regarding applying the required modifications, and expected quality control criteria from regulatory bodies (Butler and Spearman 2014). Eukaryotic hosts are the most popular recombinant protein expression systems and assist with the production of more than 60% of therapeutic proteins (Figure 2B). Mammalian cells produce 30% of therapeutic proteins produced in eukaryotic hosts. CHO (Chinese hamster ovary) cells are the most popular and used explicitly for producing more advanced therapeutics (Dumont et al. 2016). However, varying needs for producing different proteins have made many

other host cells available that each meets the requirements for the production of a specific range of proteins. Here, I review the most popular host cells for producing recombinant protein-based drugs and briefly summarize their advantages and disadvantages.

Bacteria and yeast

Depending on the target protein, E. coli and yeast can serve as great hosts for producing biologics. Indeed, the emergence of recombinant pharmaceutical protein production on the industrial scale started from insulin production in E. coli cells in 1973 (Cohen et al. 1973). After more than four decades and many developments in genetic engineering tools and bioprocessing methods, most worldwide insulin is still produced in E. coli. Collectively, E. coli serves as the host cell for producing approximately 30% (Figure 2B) of worldwide recombinant proteins, which highlights the suitability of this host for producing many proteins that do not have complicated folding or post-translational modifications in their structure (Baeshen et al. 2014; Berlec and Strukelj 2013; Dumont et al. 2016). Low cost and an efficient culturing process combined with the simplicity of genetic engineering via well-developed toolboxes make this host a powerful cell factory for producing a wide range of products (Mattanovich et al. 2012; Martínez et al. 2012).

As a eukaryotic host, yeast cells benefit from many of the same advantages as *E. coli* and possess the ability to perform post-translational modifications and folding steps for many products at a level that meets the quality control criteria (Martínez et al. 2012). PTMs play an essential role in the interaction of therapeutic proteins with their target, and a lack of proper PTMs, specifically, glycosylation can affect protein stability (Duan and Walther 2015). In addition, yeast cells benefit from both the efficient and fast culturing process and mimic human PTMs at an acceptable level for a group of products (Martínez et al. 2012). In most cases, if producing a protein in yeast cells satisfies quality control criteria, production of the recombinant protein would be more cost-effective than the production of the same product in other host cells, including mammalian cells (Martínez et al. 2012). However, despite a vast and growing number of research studies for improving yeast cells to perform more similar PTMs to human cells (Vieira Gomes et al. 2018; Huertas and Michán 2019; Šoštarić and van Noort 2021; Thak et al. 2020; Weis, Hartner, and Glieder 2006), there are still many proteins that suffer from poor quality when produced in yeast.

CHO cells

The natural ability of mammalian expression systems to perform the required modifications on proteins with high similarity to those in human cells makes these expression systems the preferred host for recombinant therapeutic protein production (Figure 2B). Among mammalian expression systems, CHO cells account for producing approximately 70% of therapeutic proteins, which indicates the importance and applicability of this cell factory in the current niche of the therapeutic protein production industry. The advantages of CHO cells over other mammalian hosts are summarized below:

I. High specific productivity (q) in CHO cells compared to other mammalian cells (Kim, Kim, and Lee 2012), as well as the high viable cell concentration in the cell culture process, which leads to a higher biomass of the cells (O'Callaghan and James 2008).

II. CHO cells have shown good genome stability (Worton, Ho, and Duff 1977; Malm et al. 2020).

III. Long years of using CHO cells have established this group of cell lines as a well-known and safe cell factory for regulatory institutions and has resulted in faster approval of new proteins produced in CHO cells (Kim, Kim, and Lee 2012).

IV. CHO cells can produce proteins with high similarity to natural human proteins (Kim, Kim, and Lee 2012). However, CHO cells have also shown a failure to produce some specific proteins: called difficult to express proteins (Tegel et al. 2020).

V. CHO cells have the advantage of easy adaptation to growth in serum-free culture media, which is preferred from a regulatory perspective. The ability to grow in suspension conditions in large-scale bioreactors is one of the fundamental requirements that a host cell factory needs to be used on an industrial scale (Xing et al. 2009).

Besides the different advantages of CHO cells listed above, the failure to properly fold some recombinant proteins (Le Fourn et al. 2014) and the observed deviation from natural human PTMs in some cases (Jenkins 2007) has motivated recent studies for (I) improving CHO cells to perform the required folding and post-translational modifications on desired proteins (Jenkins 2007; Jenkins, Murphy, and Tyther 2008; S. Fischer, Handrick, and Otte 2015; Le Fourn et al. 2014) or (II) developing alternative cell factories that could naturally perform folding and post-translational modifications with high similarity to those occurring in human cells, which are also able to produce proteins with high specific productivity (Dumont et al. 2016).

HEK293 cells

In a recent large screening study (Tegel et al. 2020), it was shown that out of the 2,189 predicted human secretory proteins, CHO cells could produce 1,276 proteins (58%) without additional protein-specific optimization. The remaining proteins failed due to different reasons such as aggregation, degradation, and folding problems. Specifically, the failure in the production of 126 of these proteins was identified as protein degradation. By switching their production in HEK293 cells, it was possible to produce 86 of the 126 proteins (68%), suggesting that HEK293 cells could be considered a competent alternative for producing difficult to express proteins in CHO cells.

HEK293 cells are the most popular human-derived cell factory used to produce proteins on an industrial scale. Besides the ability of HEK293 cells to produce a good number of difficult to express proteins in CHO cells, there are some additional advantages HEK293 cells have over other mammalian hosts:

I. High growth rate and cell density are comparable with CHO cells (Schwarz et al. 2020; Liste-Calleja, Lecina, and Cairó 2013).

II. The flexibility of HEK293 cell metabolism supports growth in different conditions (Román et al. 2016; Saghaleyni et al. 2020).

III. HEK293 cells can make an identical pattern of PTMs with natural PTMs in human cells (Goh and Ng 2018). However, this feature depends on the recombinant proteins of interest (Böhm et al. 2015).

Despite the advantages of HEK293 cells, the lower specific productivity of this cell-line compared to CHO cells has lowered the popularity of these cells among protein manufacturing companies. Also, its genome instability has raised concerns among regulatory institutions because HEK293 cells are not as well studied or established as CHO cells in the context of recombinant protein production (Y.-C. Lin et al. 2014).

Finally, to leverage the natural ability of HEK293 cells to perform human-consistent PTMs and proper folding of recombinant proteins, and to consider this cell-line as an adequate alternative to CHO cells, two main challenges need to be addressed: (I) improving our knowledge about the differences in the genome of different HEK293 cell types to better understanding how their genome instability could affect the cell phenotype and gene expression profile (II) further development of HEK293 cells from a recombinant protein production perspective to improve the specific productivity of the recombinant proteins.

Deficiency in the protein folding process leads to proteopathy diseases

Another benefit of studying the protein secretion process is to understand better its involvement in diseases where a deficiency in the production of one or a group of proteins causes a systemic failure in the functionality of a group of cells in a tissue (Figure 3) (Dugger and Dickson 2017; Jucker and Walker 2013). A failure to produce some proteins could cause a structural or functional abnormality within the cell and lead to a wide range of disorders (L. C. Walker and LeVine 2012). Diseases that are caused by such abnormalities are called proteopathies. Depending on the misfolding and tissue of origin, proteopathic diseases could lead to different cell responses and different symptoms at the organismal level (Levenson, Sturm, and Haase 2014). For example, the abnormal folding and deposition of proteins in the brain is the potential cause of a variety of neurodegenerative diseases, including dementia spectrum disorders such as Alzheimer’s disease (AD), Huntington's disease (HD), and Parkinson's disease (PD) (Price, Borchelt, and Sisodia 1993). Proteopathic disorders are not limited to neurodegenerative diseases; for example, abnormal extracellular deposition of polymeric insulin fibrils causes the emergence of a group of symptoms classified as insulin-derived amyloidosis (Gupta, Singla, and Singla 2015). There are many other examples of proteopathic diseases that repeat the same pattern of misplacing incorrectly folded proteins inside or outside of the cell and interfering with the natural processes of the cell (Luheshi, Crowther, and Dobson 2008).

The exact reason for misfolding of proteins and the generation of aggregates is uncertain in many proteopathic disorders and has remained at the level of speculation and discussing different hypotheses that may better fit the available evidence (Long and Holtzman 2019). However, it is generally understood that the level of misfolded intracellular proteins will progressively increase

with worsening symptoms in patients suffering from such diseases (Bogdanovic et al. 2020). So, based on the best of our current knowledge, the leading cause of the emergence of proteopathic diseases probably is the lack of correctly folded proteins (L. C. Walker and LeVine 2012). Hence, understanding the reasons for misfolding of problematic proteins in such diseases could probably suggest targets for therapeutic purposes (Figure 3A). However, protein misfolding could be a byproduct of some other failure in the protein production network, such as a lack of resources to fold or perform quality control. Therefore, understanding why some proteins misfold requires a systematic approach that considers the whole protein secretion pathway, of which protein folding is only one part (Figure 3B).

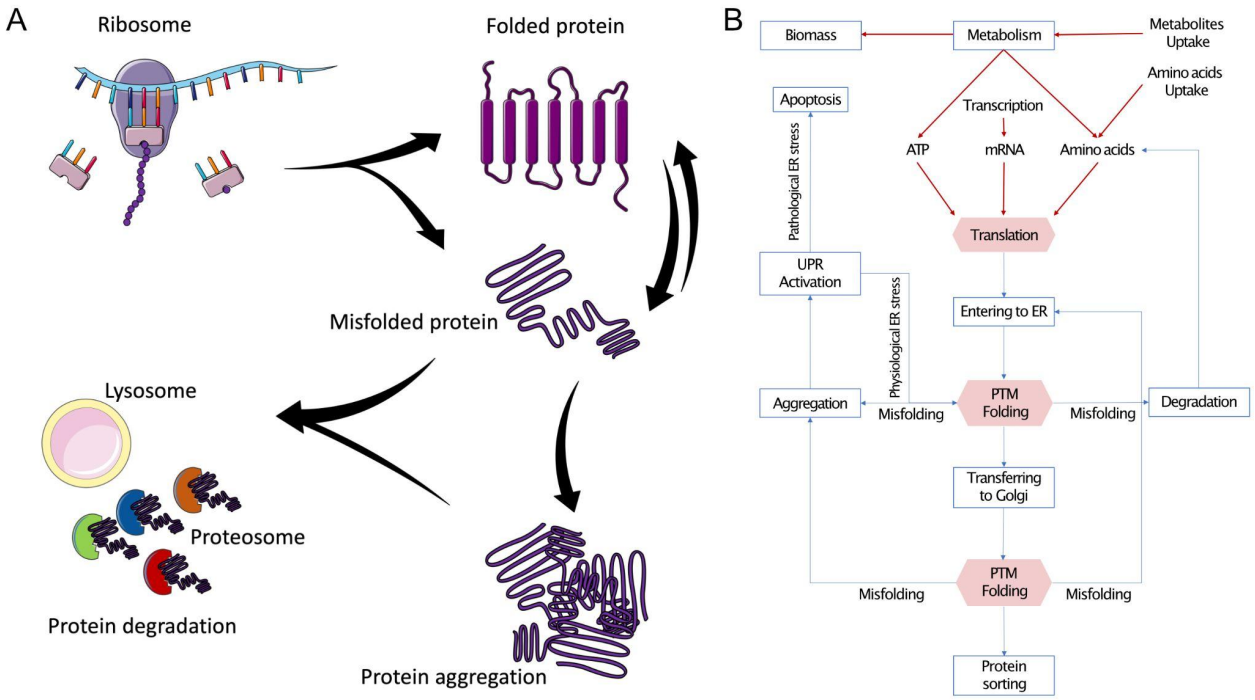

Figure 3. Deficiency in the protein folding process serves as the main reason for proteopathy diseases. (A) Aggregated misfolded proteins could cause a defect in cell metabolism and raise the risk of proteopathy disorders (B) Protein misfolding could happen because of a wide range of reasons, from metabolic deficiencies to failure in required energy for folding and protein folding machinery.

Current approaches in systems biology of protein secretion

Studying large systems requires information or logical predictions about individual components as well as the interactions between components. Analyzing the interactions between components within a system leads to a new understanding of whole-network properties that may not be possible to infer by studying each component in isolation; in the context of systems biology, these are called emergent properties of biological networks (Bhalla and Iyengar 1999; B. Palsson 2006). In this way, systems biology approaches aim to consider the organism (cell) as a whole system rather than an isolated analysis of individual network components (Tavassoly, Goldfarb, and Iyengar 2018). However, the scope of a system can vary in different contexts, from a small regulatory network with a few components to global omics analyses that consist of thousands of proteins and metabolites or even studying multi-cellular microbial communities (Voit 2013).

In a recent study, Feizi et al. (Feizi et al. 2017) showed that genes encoding the protein secretion machinery in human cells are expressed in a tissue-specific pattern to meet protein secretion demands for each specific tissue. Other studies highlight how protein secretion rate and PTM pattern are affected by the cell culturing condition as well as the status of different pathways within the cell, such as the composition of culture media, redox homeostasis, apoptotic pathways, and energy metabolism (Schwarz et al. 2019; Meuris et al. 2014; Del Val, Polizzi, and Kontoravdi 2016; Liang et al. 2020; Behrouz et al. 2020; Rahimpour et al. 2013). These studies imply that the protein secretion process, as a large pathway spanning multiple cellular compartments, has tight connections with the other signaling and metabolic pathways within the cell (Lodish et al. 2000). Hence, studying such complex processes via reductionist approaches, which only consider individual components of the network in an isolated setup, is insufficient to reach a more comprehensive understanding of the protein secretion process. Instead, holistic approaches that consider all components and their corresponding interactions within and beyond the pathway may pave the way to discover other mechanisms that govern the protein secretion process. However, to consider all components of the network in the analysis, we first need collective characterization and quantification of components within the network, and next interpret the final state of the network through analysis of collected information for each component and the interactions between components; what I call omics analysis (Micheel et al. 2012).

In recent years, three landmark achievements in the field of molecular and computational biology greatly impacted the research and assisted in developing our knowledge about the protein secretion process: (I) Availability and rapid cost decrease of global omics approaches for high throughput measurements of molecular components such as genes, proteins and metabolites (Reel et al. 2021; Lancaster et al. 2020; Sun and Hu 2016; Akiyama 2021) (II) Improvements in the algorithms for analyzing big data sets using novel comparative and deep learning approaches (Uhlén et al. 2019; Obudulu et al. 2018; Schinn et al. 2021; D. Lin et al. 2020; Robinson et al. 2019; L. Zhang et al. 2018; Reel et al. 2021; Martorell-Marugán et al. 2019) and (III) emergence of the CRISPR genome engineering method in addition to other advancements in genetic engineering tools (Sergeeva et al. 2019; Cong et al. 2013). These advancements paved the way for testing more ideas more efficiently, thereby accelerating the design-test-investigate loop (M. Zampieri et al. 2017; Parola, Neumeier, and Reddy 2018; Kweon and Kim 2018; Cupp-Sutton and Wu 2020; Toby, Fornelli, and Kelleher 2016).

Omics techniques and computational data analysis approaches

Each omics approach captures a different layer of information in the cell, such as the genome, transcriptome, proteome, or metabolome. Omics technologies generally fall into two main categories (Figure 4): (I) sequencing-based approaches like genomics and transcriptomics and (II) mass spectrometry-based approaches like proteomics and metabolomics.

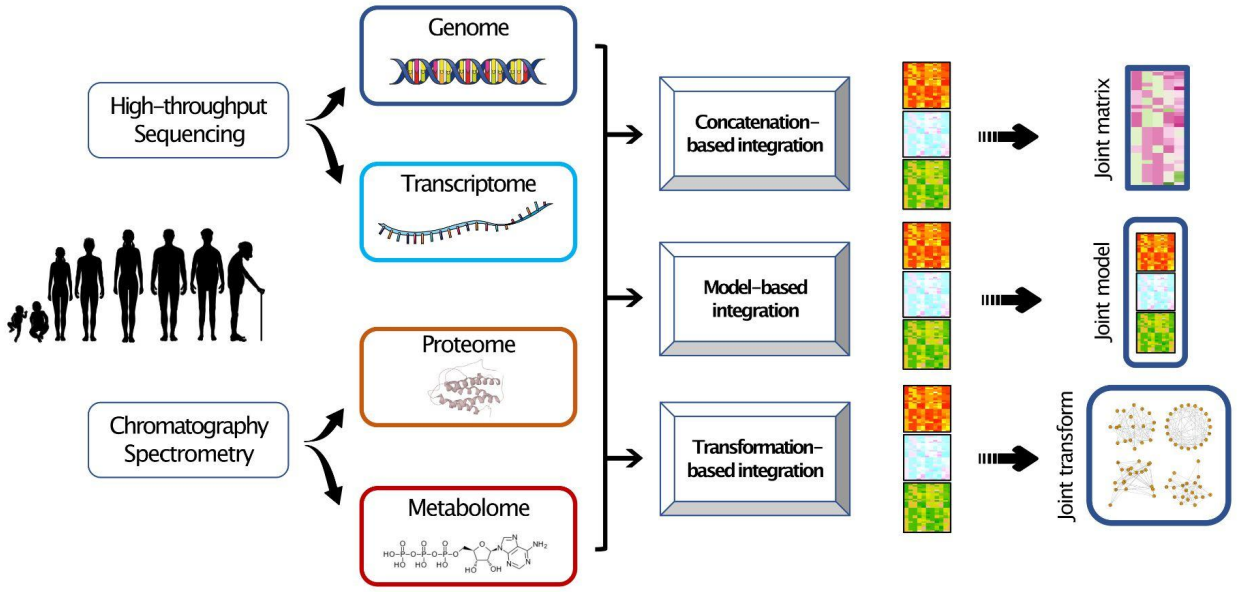

Figure 4. Methodological background and integrative analyzing disciplines for biological omics datasets.

Genomics

Genomics was the first introduced omics approach and provided access to the whole DNA sequence for many organisms. Asa field of study, DNA sequencing started right after discovering the helical structure of DNA in 1953 (Ankeny 2003). However, it took a couple more decades to complete the first whole-genome sequence for *Haemophilus influenzae*, as an individual organism, in 1995 (Fleischmann et al. 1995). Despite the availability of genome sequencing technology, until recent years and the availability of next-generation sequencing technologies, high-throughput genome sequencing was not yet market-friendly (Hall 2007; Church 2006). However, further technological development drastically decreased the costs associated with genome sequencing (Hall 2007; Church 2006). The availability of genome sequencing data for many organisms challenged extracting knowledge from big data in front of biologists for the first time. It motivated many studies that today are known as the footstones of computational biology and systems biology (McKusick 1997). For example, developing algorithms for the assembly of sequenced DNA contigs followed by functional annotation of assembled sequences; topics that are still under progress and improvements (Pevsner 2009; Pop 2009; Visser et al. 2002; Flicek et al. 2013). Depending on the research study, huge levels of information could be inferred from genomic data. For instance, comparative genomics among different organisms and cells could highlight differences from higher chromosomal structure differences (karyotype) to single nucleotide polymorphism among different alleles of a gene (Hardison 2003; Ellegren 2008).

Transcriptomics

Transcriptomics approaches also benefit from sequencing-based technologies. A transcriptome dataset consists of a set of reads belonging to RNA molecules transcribed from an organism's genome in a specific physiological condition. So, unlike the genomics profile, the transcriptomic profile of an organism is dynamic and changes with time and the physiological status of the system (Supplitt et al. 2021). In addition, the transcriptome can also capture the effect of alternative splicing and detect transcripts that belong to the same gene (J. Wang et al. 2015).

Currently, there are two leading technologies for generating transcriptome libraries: (I) microarrays, which is a relatively older technique that uses a set of pre-defined sequences for detecting reads in the samples (Chang 1983), and (II) RNA-seq, which is a newer and more common approach which benefits from high-throughput sequencing to decipher all transcripts and match them in downstream processing to a corresponding reference genome or a predefined set of transcripts (Chu and Corey 2012; Z. Wang, Gerstein, and Snyder 2009; Dobin et al. 2013; Patro, Mount, and Kingsford 2014; Wu and Watanabe 2005).

Another recent breakthrough technology in the field of molecular biology is to perform high-throughput sequencing analysis at the level of single cells (Adil et al. 2021; Aldridge and Teichmann 2020; Andrews et al. 2021). Specifically, transcriptome profiles of individual cells on a large scale enable investigation of cellular heterogeneity that is not possible when using bulk samples (Y. Wang and Navin 2015).

Currently, differential gene expression analysis is the most common approach for analyzing transcriptome data; comparative analysis of samples categorized in two or more groups and finding genes that exhibit statistically different expression levels between groups (McDermaid et al. 2019). Different approaches have been introduced for performing differential expression analysis between samples, each applying different assumptions (McDermaid et al. 2019; Quinn, Crowley, and Richardson 2018; Li 2019). Briefly, all approaches for the comparative analysis of transcriptome profiles consider the following parameters to define the proper statistical method: (I) a small number of replicates due to the laboratory limitations in producing biological samples (II) wide range of expression for a gene between different samples and also for different genes in a sample (III) presence of outliers that could be because of high difference in expression for a gene between samples or because technical noise and (IV) non-normal distribution of raw count data (Love, Huber, and Anders 2014; McCarthy, Chen, and Smyth 2012). The final result of differential expression analysis is generally a table of gene expression fold changes and the levels of statistical significance associated with the fold changes.

Mass spectrometry-based omics approaches: proteomics and metabolomics

Unlike sequencing-based approaches, global untargeted metabolomics and proteomics, which measure different classes of biomolecules in a sample, are both based on mass spectrometry (MS) approaches (Blum, Mousavi, and Emili 2018). MS is used to determine the molecular weight of different ions by measuring the mass-to-charge (m/z) ratios (Parker, Warren, and Mocanu 2011). The molecules in a sample must, therefore, first be converted to the gas phase. Then, a mass analyzer component separates ionized species based on their difference in m/z ratio. Finally, an ion detector detects and generates a signal proportional to the m/z ratio for each molecule (Parker, Warren, and Mocanu 2011). In omics analysis, the first MS run often continues with additional MS rounds on each of the detected analytes to further investigate the level of each analyte in the samples (called tandem MS or MS/MS). For this purpose, before MS measurements, fragmented peptides must first be labeled for each sample. Then, by running additional rounds of MS, one can detect the ratio for each ion between two samples in a group of samples (Di Girolamo et al. 2013). Due to the complexity of biological samples, one suggested additional preprocessing step is the fractionation of samples and decreasing the number of ions in

each fraction. For example, using liquid chromatography (LC), each sample can be separated into more fractions. Then by performing MS/MS on each fraction, a higher number of molecular species can be detected. LC-MS/MS is currently one of the standard approaches for performing proteomics and metabolomics analysis on biological samples (Di Girolamo et al. 2013). Finally, by summing up the measured intensity ratios for all species belonging to a known molecule, the final intensity ratio table for all detected proteins or metabolites can be calculated.

The final result of LC-MS/MS analysis is a table of ratios for each protein or metabolite. Ratios indicate the level of detected intensity for each protein or metabolite between a case sample and a reference sample (Taverna and Gaspari 2021), which usually serves as a pool of all samples in an experiment as a baseline. The bioinformatic approaches for analyzing the raw results of proteomics and metabolomics are not as standardized as those for transcriptomics and genomics analysis. However, comparative analysis for finding molecules with a statistically significant pattern of change between samples, followed by enrichment analysis for finding enriched pathways with significantly different proteins or metabolites, can help to reveal the pattern of alterations in metabolism and physiology between two samples.

Integrative omics analysis

Each type of omics data is like a cross-sectional layer of information from the system under study (Subramanian et al. 2020; Das et al. 2020) (Figure 4). Although each layer profoundly improves our knowledge regarding similarities and differences between samples in an experiment, only specific types of biological information emerge in each single omics analysis. Thus, a further advantage in multiple omics analysis studies could be acquiring higher levels of biological insight by applying integrative omics approaches that bridge remarkable findings between individual omics layers and fill the gap in our knowledge between observed genotype and phenotype of cells.

The main challenges in integrative omics analysis are as follows: (I) The coverage of detected molecular species may vary in different mass spectrometry analyses, which causes heterogeneous data coverage for different samples. The same issue exists in transcriptomics analysis as well, specifically newer approaches for single-cell analysis (Martorell-Marugán et al. 2019) (II) Each omics dataset has its specific formalism and needs proper data treatments like normalization and managing missing values before entering to omics integration pipelines (Mertens 2017; Nusinow and Gygi 2020; Misra et al. 2018). (III) The variance of measurements is different between omics datasets (Boccard and Rudaz 2016). For instance, in many transcriptomics studies, the absolute logarithmic fold change higher than 1 ($|Log_2FC| > 1$) is considered a cut-off for DE genes. In contrast, in proteomics datasets, the variance of changes is much lower than transcriptome, and there is no agreed cutoff for fold change in research with proteomics analysis. The different dynamic ranges between different omics layers is thus an important consideration when performing an integrative analysis. (IV) In biological studies, the number of measured features (genes, proteins, metabolites, etc.) is often more than the number of samples. This causes a high-dimensional space with many correlated features that may complicate or interfere with many downstream analyses (Subramanian et al. 2020). To overcome this challenge, dimension

reduction methods such as PCA (principal component analysis) (Jolliffe and Cadima 2016) and LDA (linear discriminant analysis) (Tharwat et al. 2017) can be used.

Omics data integration can be performed sequentially or simultaneously (Bersanelli et al. 2016; Subramanian et al. 2020). In the former, algorithms perform tests on omics layers in sequential order. Significant alterations in each omics layer in agreement with findings on the previous layers will be selected to detect modules changing across all omics layers between two sample groups. On the other hand, simultaneous multi-omics integration approaches secure high sensitivity in considering all relationships between variables but minimize the use of existing knowledge about variables in different omics layers (Silverbush et al. 2019).

Another approach to categorize multiple omics integration methods is based on their mathematical aspects (S. Huang, Chaudhary, and Garmire 2017). From a mathematical perspective, omics integration methods can be classified into three different groups (Figure 4): (I) concatenation-based integration methods, (II) model-based integration methods, and (III) transformation-based integration methods (Reel et al. 2021). In addition, the level of supervision could vary from supervised to semi-supervised to unsupervised and depends on the specific approach ("Handbook of Statistics" n.d.; Stein-O'Brien et al. 2018; Wilson et al. 2019; Bersanelli et al. 2016).

Concatenation-based integration

Concatenation-based integration approaches simply combine matrices from different omics datasets to generate a joint matrix. Once the joint matrix is generated, it will be used for downstream supervised or unsupervised analysis (Reel et al. 2021). Because of the high number of variables in the joint matrix, a feature selection step is usually necessary before any downstream machine learning analysis (Sorzano, Vargas, and Pascual Montano 2014). The reduced joint matrix could be used as input for different supervised (e.g., decision tree, artificial neural network, random forest, etc.) (Quinlan 1993; Domingos and Pazzani 1997; Breiman 2001) or unsupervised approaches (e.g., non-negative matrix factorization, iCluster, iCluster+, MoCluster, MOFA, etc.) (S. Zhang et al. 2012; Shen, Olshen, and Ladanyi 2009; Mo et al. 2013; C. Meng et al. 2016; Argelaguet et al. 2018). Matrix factorization approaches attempt to reduce dimensionality by inferring a different (smaller) set of variables from the data (Pierre-Jean et al. 2020). The most well-known approach of this type is principal component analysis (PCA). Canonical correlation analysis (CCA) is another approach that investigates the relationship between variables in two or more datasets (D. S. Wilks 2011). This approach has been widely used in omics integration analysis (Rodosthenous, Shahrezaei, and Evangelou 2020). In general, concatenation-based approaches need proper preprocessing regarding normalization before concatenation. A disadvantage of these approaches is that they do not consider the potential differences in distributions among different types of omics datasets. Also, handling a large matrix of variables from different omics datasets can become computationally expensive.

Model-based integration

In model-based integration approaches, first, multiple intermediate models simulate each of the omics datasets, and then a final model integrates the intermediate models (Reel et al. 2021). Since each dataset is first modeled individually, this approach effectively addresses the challenge

of combining omics datasets because of their methodological differences, such as the variance of reported values for measured variables and differences in absolute or relative measurements. Then ML methods can be used to train a general joint model that selects the most important variables from each model (Drăghici and Potter 2003). Similar to concatenation-based integration, here, the level of supervision may vary between the different methods. An advantage of model-based integration approaches is their flexibility in analyzing omics data from different sets of samples. However, these could be a disadvantage because if omics data are highly heterogeneous, medium to weak signals could easily be lost (López de Maturana et al. 2019; Kamisoglu et al. 2017).

Transformation-based integration

In transformation-based integration approaches, each omics dataset is first transformed to a graph or kernel matrix. Then a joint matrix will be constructed from these matrices (Reel et al. 2021) to train a machine learning model. These approaches can organize multiple datasets that measure different variables for a fixed group of samples (Yan, Zhao, and Pang 2017). Also, the computational cost for these approaches is lower compared to the other two groups.

Modeling approaches

Individual and integrative omics analyses can provide a great deal of molecular information about a biological system. However, each omics sample only captures the characteristics of a biological system at a single time point. Therefore, it is beneficial to use predictive tools to contextualize data obtained from high-throughput experiments and accelerate the design-test-investigate loop (Nielsen 2017a).

Different approaches exist for modeling the reactions and interactions in a cell and can generally be categorized as topological, stoichiometric, and kinetic models (Steuer 2007). Topological models consider the directions and interactions between components in a network. So these models are mainly limited to predicting global qualitative properties of the network, such as modules and hub nodes (Najafi et al. 2014). Unlike topological models, kinetic models predict cell states with quantitative measurements; however, developing kinetic models requires a lot of experimental data to assign or estimate the many parameter values, which are most often difficult to obtain or unavailable (B. Ø. Palsson 2011). On the other hand, metabolic models can provide quantitative predictions and are flexible for contextualizing experimental data if such data are available (Nielsen and Hohmann 2017; B. Ø. Palsson 2015). Specifically, recent advances in integrating molecular data with metabolic models have paved the way for using proteomics and metabolomics data as additional constraints in genome-scale metabolic models (Domenzain et al. 2021; Pandey, Hadadi, and Hatzimanikatis 2019; Yu Chen, Nielsen, and Kerkhoven 2021; Xia et al. 2021; Yu Chen and Nielsen 2021). In the present thesis, genome-scale metabolic models were used to study cell metabolism.

Genome-scale metabolic models

A genome-scale metabolic model (GEM) is a mathematical representation of a set of biochemical reactions in an organism of interest (Nielsen 2017b). Reactions are single units of the model that consume and produce metabolites and can be associated with one or more genes.

Principles

Information about reactions and metabolites comprising the metabolic network of a cell or tissue is organized in a M × N stoichiometric coefficients (*S*) matrix (B. Palsson 2006), where M is the number of metabolites and N represents the number of reactions (Figure 5A-B). The S matrix's negative or positive values correspond to consumption or production of metabolites in their corresponding reactions, respectively. Besides reactions annotated from the genome of organisms and indexed in metabolic databases such as KEGG (Kanehisa et al. 2021) and METACYC (Caspi et al. 2008), other reactions exist in the *S* matrix; these reactions mainly have one of the following roles within the genome-scale models:

I. Lump reactions: these reactions are used for connecting a group of metabolites. The most commonly used lump reaction in GEMs is the biomass reaction, which consumes metabolites that comprise the cell biomass. Stoichiometric coefficients in the biomass reaction are assigned based on experimental analysis of the dry cell composition or based on estimations from biomass of organisms phylogenetically close to the organism of interest (Schulz et al. 2021; Mendoza et al. 2019).

II. Exchange reactions: these reactions enable the exchange of metabolites into and out of the system (Figure 5A-B).

III. Transport reactions: reactions that transport metabolites from one model compartment to another (e.g., from the extracellular environment to the cytoplasm) (Figure 5A-B).

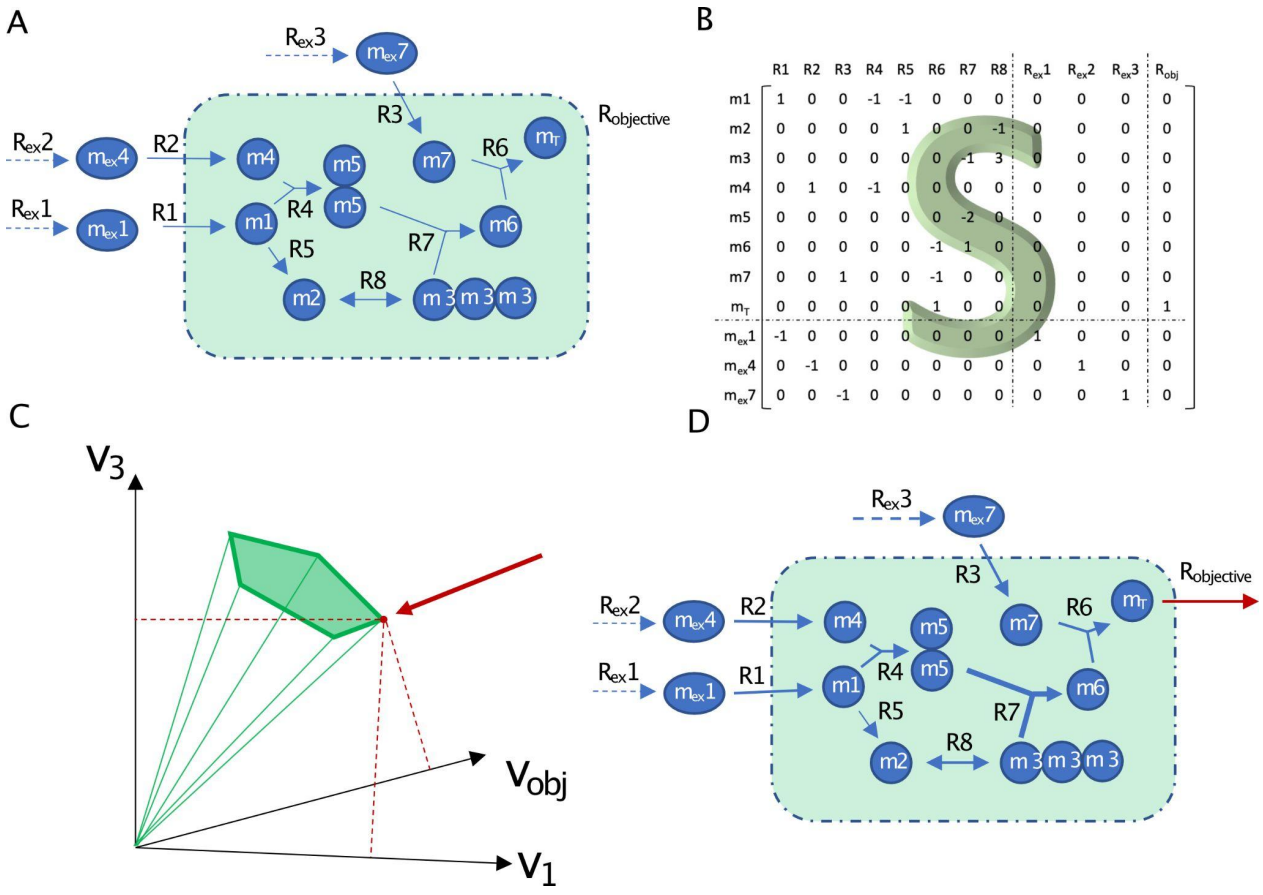

Figure 5. Principles of genome-scale metabolic modeling and flux balance analysis (FBA). (A) Visualized form of a toy metabolic model including exchange reactions, R_{ex}, transport reactions, R_1, R_2, and R_3, an objective function reaction, $R_{objective}$, and intracellular reactions R_4-R_8. m is used for indexing metabolites (note the number of consumed and produced metabolites in each reaction) (B) Stoichiometric matrix (*S*) representing the toy model in (A). Metabolites are in rows of the matrix and reactions as columns. Numbers in the stoichiometric matrix correspond to the number of consumed (negative) or produced (positive) equivalents of each metabolite in each reaction. (C) Visualized flux cone based on the fluxes of three reactions (R_1, R_3, and R_{obj}) in the model. (D) Visualized fluxes of reactions based on the results of FBA analysis. Arrow width represents the flux rate for each reaction.

Compared to other cellular processes such as translation, cell cycle, and replication, metabolism is a relatively fast process. Therefore, when simulating reaction fluxes using flux balance analysis (FBA) (Rajvanshi and Venkatesh 2013), it is assumed that a steady-state condition applies, meaning that the consumption and production for each metabolite must be equal (except exchange metabolites and biomass) (Nielsen and Hohmann 2017; B. Ø. Palsson 2015). Using the steady-state assumption and applying other constraints for the consumption or production of specific metabolites, estimating the flux distribution for all reactions for that system and condition is possible. However, to calculate the flux distribution, additional assumptions are required:

I. The minimum and maximum permitted flux for each reaction in the model are lower-bound (lb) and upper-bound (ub). Thus, the directionality of reactions can be defined using lb and ub values, where an irreversible reaction has a lb of zero.

II. Changes in extracellular metabolite concentrations over time can be measured and converted to flux values, which can then constrain the uptake or production of those metabolites in the model.

The flux range for all reactions together comprises the feasible flux space that satisfies all the mass balance and thermodynamic constraints (Nielsen and Hohmann 2017; B. Ø. Palsson 2015) (Figure 5C). This is also called the solution space or flux cone, and each point in this theoretical space represents a flux vector (*v*), which defines the flux of each reaction ("Metabolic Flux Analysis" n.d.). Depending on the study, the optimal solution could be defined as the point in the flux cone that minimizes (or maximizes) an objective function. This method is known as flux balance analysis (FBA). For example, the biomass reaction flux is often defined as the objective function in FBA, assuming that cells (specifically microbial) tend to maximize growth. However, any reaction or combination of reactions can be used as an objective function. For instance, minimizing the consumption of ATP, maximizing secretion of a specific metabolite for metabolic engineering purposes, or minimizing lactate production (Yiqun Chen et al. 2019; Baart and Martens 2012; Nilsson and Nielsen 2016; Nilsson et al. 2020; Feizi et al. 2013; Gutierrez et al. 2020) are potential objective functions.

Reconstruction

The process of reconstructing a GEM starts from the genome as a library of genes present in an organism, which are then associated with the reaction(s) that they are known to catalyze. These

approaches are called bottom-up reconstructions (Thiele and Palsson 2010). Next, by adding exchange and transport reactions and introducing a biomass reaction, the reactions in the model can potentially carry flux. However, if all metabolites in the reactions are not connected, it may lead to gaps in the model and a need for performing gap-filling steps (Fouladiha et al. 2021). This further complicates the laborious process of reconstructing functional genome-scale metabolic network models.

To overcome these challenges and accelerate reconstructing GEMs, semi- or fully automatic approaches have been under development in recent years, aiming to reconstruct a cell-specific model from various inputs and automate gap-filling steps (Mendoza et al. 2019). Besides improving the efficiency of GEM reconstruction by such approaches, the quality of reconstructed models in terms of accuracy of predictions and also reproducibility of the process (Opdam et al. 2017) also increases by using (semi-)automated approaches. In this thesis, I have used the RAVEN toolbox and INIT algorithm to generate genome-scale metabolic models starting from transcriptomics data and the generic human GEM, Human1 (Agren et al. 2012; H. Wang et al. 2018; Agren et al. 2014). The INIT is an algorithm that starts from transcriptomics data instead of the genome to reconstruct specific models based on expressed genes in a transcriptome profile (Agren et al. 2012).

Human1

Constraint-based modeling of human metabolism has been an active area of interest over the past couple of decades, where different GEMs have been developed to model human metabolism (Duarte et al. 2007; H. Ma et al. 2007; Thiele et al. 2013; Brunk et al. 2018; Mardinoglu et al. 2013). The most recent GEM introduced for human cells is Human1 (Robinson et al. 2020). Human1 integrates all curated data deposited in the previous human metabolic models, has included higher quality gene protein reaction (GPR) associations, and presents a web portal hosting the GEM content that enables the overlay of gene or protein expression data on the pathway- or compartment-level maps.

Further GEM improvements

Since the first developed genome-scale metabolic model in 1999 (Edwards and Palsson 1999), generating a GEM for organisms is much more efficient and accessible (Gu et al. 2019). Improving the approaches and algorithms to generate better GEMs with improved power of prediction and higher coverage of pathways within the cell has always been an active area of interest (Mendoza et al. 2019; Opdam et al. 2017; Pfau, Pacheco, and Sauter 2016; G. Zampieri et al. 2019). Efforts to improve GEMs could be categorized as two main categories that both aim to contextualize information from high-throughput analyses, such as metabolomics, proteomics, and phosphoproteomics (Figure 6) (Hadadi et al. 2020; Töpfer, Kleessen, and Nikoloski 2015; Hastings et al. 2019; Töpfer, Seaver, and Aharoni 2018):

I. Improving model predictions by logically constraining reactions and further reducing the flux cone solution space (Lloyd et al. 2018; Lewis, Nagarajan, and Palsson 2012; Bordbar et al. 2014; O ’brien et al. 2013; Lerman et al. 2012; Oftadeh et al. 2021; Sánchez et al. 2017). For example, Sanchez et al. (Sánchez et al. 2017) provided an algorithm called

GECKO, wherein users can constrain the upper bound of reactions using absolute protein concentrations (from proteomics analysis) and specific reaction rate constant (kcat) for each protein. The resulting model is called an enzyme-constrained GEM (ecGEM) (Figure 6A).

II. Improving the coverage of reactions in the model by generating specific reactions for subcellular pathways that are not cataloged in the classical GEMs, for instance, translation, lipid production, and protein secretion pathways (Nookaew et al. 2008; Du et al. 2019; Oftadeh et al. 2021; Loira et al. 2012; Feizi et al. 2013; Gutierrez et al. 2020).

Each of the mentioned approaches has dramatically benefited from the flexibility of GEMs to add recently discovered biological knowledge, followed by significant improvements in the power of predictions by the model. In addition, the protein secretion process has also been added to classical GEMs in multiple studies (Figure 6B).

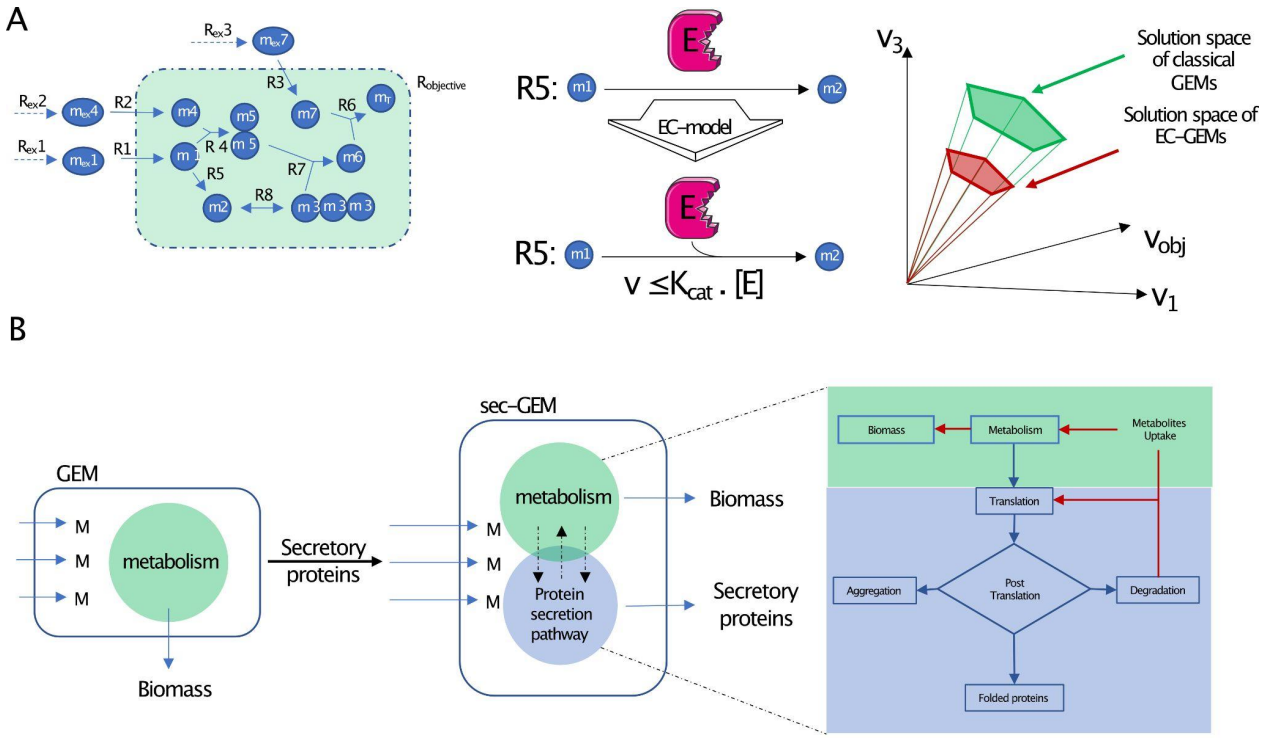

Figure 6. Genome-scale metabolic model (GEM) improvements. (A) Improving GEM predictions by applying constraints on reaction fluxes based on proteomics data further limits the solution space. The resulting model is called an enzyme constrained GEM (EC-GEM). (B) Improving GEMs coverage by expanding the model to cover reactions and pathways within the cell that are not present in classical GEMs; for example, the protein secretion process (sec-GEM).

Protein secretion modeling

Modeling of protein secretion processes mainly fall into two groups: (I) modeling the pathways and reactions that are in charge of translation, folding, and sorting of proteins (Feizi et al. 2013; Gutierrez et al. 2020; Thiele et al. 2009) and (II) developing models that predict location and structure of post-translational modifications (PTMs) for each protein aiming for model-derived glycoengineering (Stach et al. 2019; Spahn et al. 2016; Štor et al. 2021). In the latter approach,

the production process of proteins is not the topic of study. Instead, only discovering and modeling the logic for producing desired patterns of PTMs is of interest.

Feizi et al. (Feizi et al. 2013) developed the first genome-scale protein secretion model (secGEM). They began with yeast GEM and adding protein secretion pathway reactions to yield a functional model capable of producing all detected secretory proteins in yeast. However, each secretory protein has specific characteristics, and all these proteins are not possible with one set of reactions. To overcome this challenge, they developed a pipeline that generates a particular set of reactions based on characteristics for each protein and adds these protein-specific reactions to the reference model. By applying such an approach, they demonstrated that it was possible to systematically predict the metabolic demand for each secretory protein's production and quantitatively estimate each component of the secretory machinery.

More recently, Gutierrez et al. in 2020 (Gutierrez et al. 2020) applied the same approach to reconstruct secGEMs for Chinese hamster ovary (CHO) cells, human cells (using Recon 2.2 human GEM as reference model), and mice. The model successfully simulated the energetic costs and machinery demand for the production of each secretory protein. In addition, simulations suggested alterations in the expression of secretory proteins in various conditions that were later confirmed by experimental analysis.

In the modeling section of the current thesis, by taking a similar approach, I developed a toolbox to expand Human1 to cover protein secretion reactions that are capable of producing all secretory proteins detected via high-throughput proteomics profiling.

Aim and significance

In this thesis:

In **paper I**, I compared the most popular HEK293 cell lines currently widely used in different laboratories for protein production. In addition, I compared genome and transcriptome between different clones and discovered genomic and transcriptomic alterations that could lead to growth in the suspension condition.

In **paper II**, I investigated the differences in HEK293 cell metabolism and physiology due to the production of secretory or non-secretory recombinant proteins. I further explored genes whose expression correlated with recombinant protein production. I identified a list of genes that exhibited a significant expression change pattern between different erythropoietin (EPO) production levels.

In **paper III**, I continued analyzing secretory recombinant protein production by generating an omics dataset including transcriptome, proteome, and metabolome for three HEK293F cell lines that produce EPO at different rates. Using comparative and integrative analyses, I sought to identify metabolic and signaling patterns that correlate with protein production. Furthermore, I took a modeling approach and developed a toolbox for constraint-based modeling of the protein secretion process. I used to find host cell proteins (HCP) that showed the highest competition with EPO in resource usage. These proteins constituted knock-out candidates with the potential to improve EPO production.

In **paper IV**, I investigated changes in the expression patterns of proteins involved in the secretion process in different cancer types. I performed a comparative analysis and implemented a machine learning approach to analyze gene expression data from different cancer types. As a result, I identified critical genes with a clear pattern of alteration between normal and tumor samples and between samples from different cancer stages.

Results & discussion

Key genes in cell morphology transformation (Paper I)

To study the protein secretion process in human cells, I chose HEK293 cells as a model cell line and first aimed to find differences between HEK293 cell lineages. HEK293 cells initially originated by immortalizing human embryonic kidney cells of an aborted female embryo transformed by integrating a four kbp adenoviral 5 (Ad5) genome fragment (Graham et al. 1977; Louis, Evelegh, and Graham 1997). HEK293 cells are the most common human cell-line for the production of recombinant proteins in a wide range of research (Dumont et al. 2016). To improve recombinant protein production, HEK293 cells have been under development for a long period, which has led to the establishment of several HEK293 cell lineages (Dumont et al. 2016; Goh and Ng 2018). In addition, various cell-line engineering strategies and approaches have resulted in the availability of different adherent and suspension cell lines (Figure 7A) (Malm et al., n.d.; Graham 1987; Garnier et al. 1994; Côté et al. 1998; Schwarz et al. 2020).

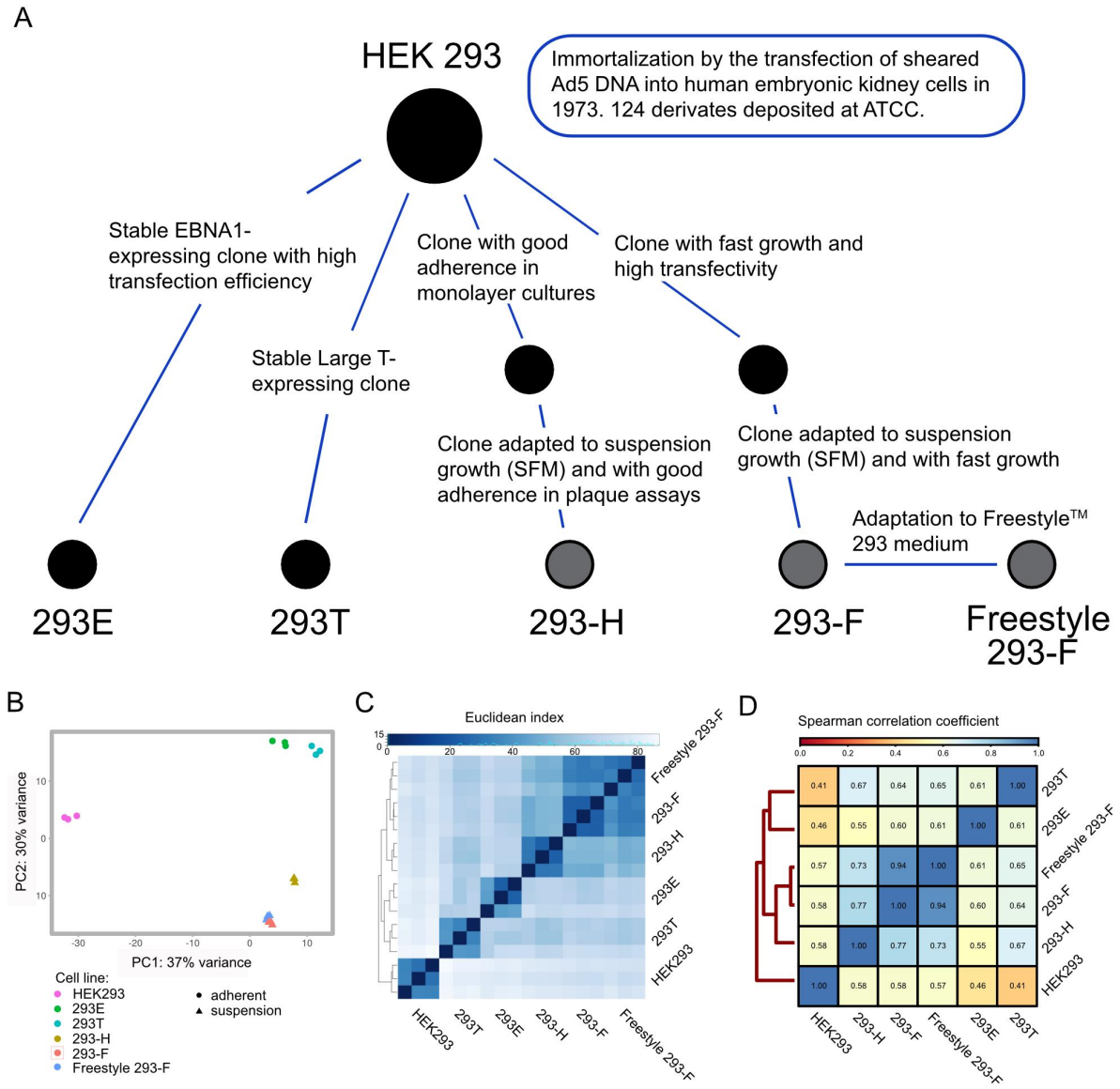

Figure 7. Overview of the six HEK293 lineages investigated in our study and their relationships. (A) Genomic and transcriptomic comparisons of HEK293 cells showed a taxonomic divergence between parental HEK293 and progeny cell lines. Black dots represent adherent cells, and gray dots correspond to suspension cell lines. (B) PCA of HEK293 cells using the transcriptome shows the separation of progeny cell-lines and parental cells in the first component of PCA (C) Hierarchical clustering of HEK293 cells by calculating the Euclidean distance between transcriptomic profiles of different clones (D) Genomic comparison of HEK293 cells based on the Spearman correlation coefficient of the read counts.

Genomic differences between HEK293 parental and progeny cell lines

Host cell factories in general, and HEK293 cells specifically, exhibit genome instability (Stepanenko and Dmitrenko 2015; Wurm 2013; Vcelar et al. 2018). Long-term cultivation and subcloning of cells result in genomic alterations between subclones. To investigate the differences between HEK293 clones, six industrially relevant HEK293 cell lines were cultured (Figure 7), followed by sample collection for genomics and transcriptomics analysis. HEK293 cell lines in our study were either adherent (HEK293 parental cell, 293T and 293E) or suspension (293-F and Freestyle 293-F, 293-H). Hierarchical clustering of transcriptome and genome divided cell lines based on their morphology into two groups (Figure 7C-D). However, PCA results showed separation between progeny and parental cell lines in component 1 (Figure 7B), suggesting genomic divergence between progeny and parental cell lines than among different progeny lines.

I analyzed genome and transcriptome datasets to investigate differences between HEK293 clones. Genome comparison between clones highlighted a group of genes with the same copy number gain or loss in the chromosomes of progeny cells compared to parental HEK293 cells (Figure 8). On chromosome 13 of all progeny cells, a region of > 15 MB was amplified. Seven protein-coding genes detected in this region exhibited at least a 2-fold increase in the number of copies in progeny cells. Four out of seven genes (BORA, MZT1, PIBF1, and KLHL1) belonged to the cytoskeleton gene set. Among other genes with a different copy number between clones, fumarate hydrolase (FH) showed a copy number loss in some clones (Figure 8 - chromosome 1). In our analysis, I detected the loss of gene copy numbers for FH and also its neighboring gene kynurenine 3-monooxygenase (KMO) in 293F, Freestyle 293F, and 293E clones (log2-fold copy ratio of <-1). The KMO gene has been previously reported for losing gene copies in HEK293 cells and is hypothesized to play a role in the transformation of HEK293 cells (Y.-C. Lin et al. 2014).

Following the observed pattern of common genomic changes between progeny cells and the parental cell line, I identified SNPs among the genomes of all progeny cell lines compared to the parental HEK293 cell line. I performed enrichment analysis on common genes with moderate to high SNP impact. I found significant (adjusted p-value <0.05) enrichment of homophilic cell adhesion via plasma membrane adhesion molecules (adjusted p-value 0.025) and cell-cell adhesion via plasma-membrane adhesion molecules (adjusted p-value 0.032). In addition, a group of significantly differentially expressed genes enriched in the protocadherins protein family (PCDH12, PCDHB10, PCDHB13, PCDHB15, PCDHB16 PCDHGA2, PCDHGA3, and PCDHGB2). Although the exact impact of SNPs on these proteins is not well known, enrichment

of common SNPs within these proteins and between all progeny HEK293 cells may suggest a specific biological role.

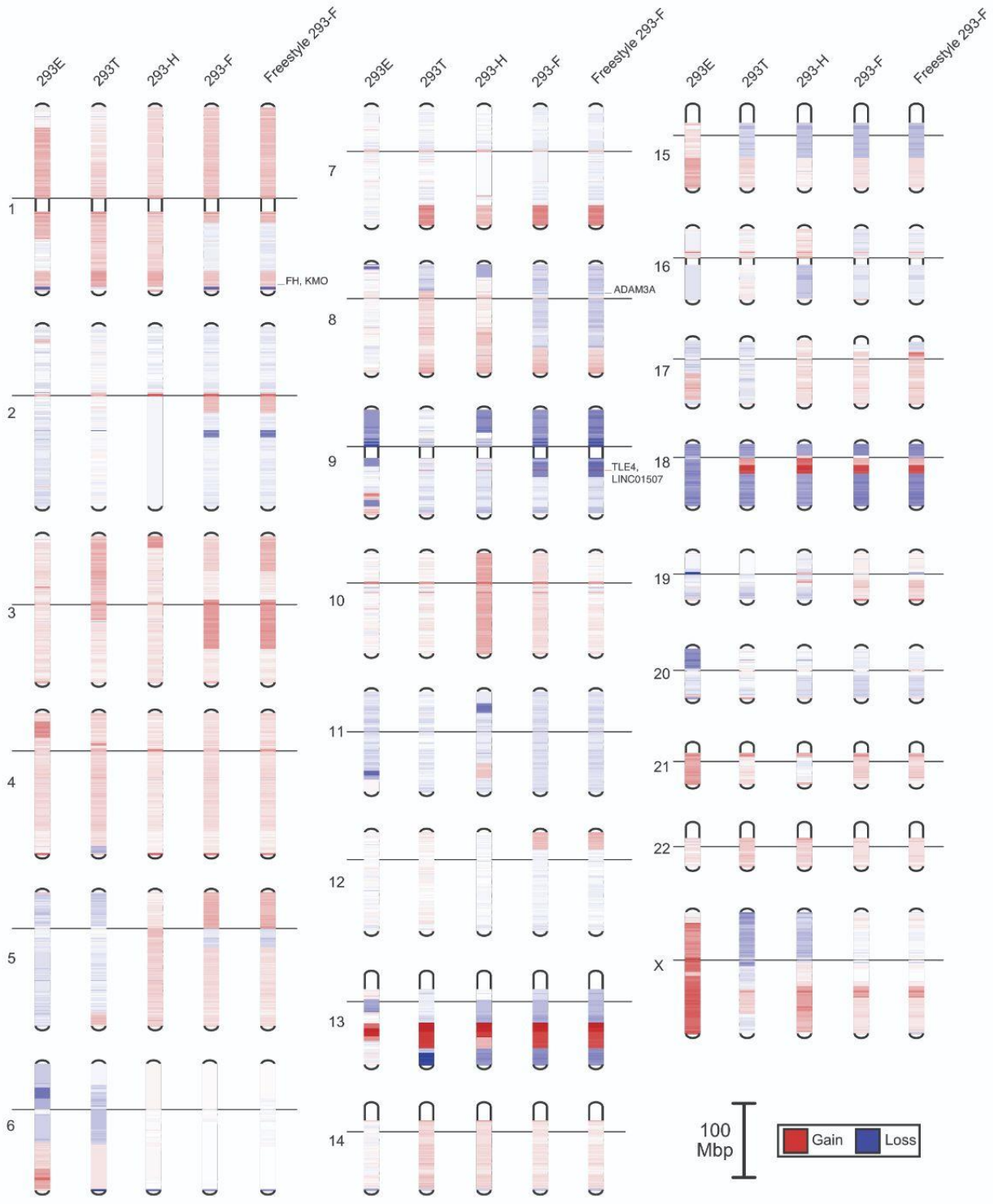

Figure 8. Genomic differences between HEK293 progeny cells and the HEK293 parental cell line. Red color shows gain, and blue indicates loss over progeny cells compared to parental clones. Gene names indicate genes that showed the same pattern of gain or loss between all progeny cells compared to the parental cell line.

Transcriptomic comparison of HEK293 cells

Pairwise comparison of HEK293 clone transcriptomes showed that the parental clone exhibited the greatest difference among the cell lines regarding the number of differentially expressed genes (Figure 9). I detected 329 genes with a constant and significant (Benjamini-Hochberg adj. p-value < 0.05, absolute log2 fold change > 1) pattern of change, which have different expression patterns between all progeny cells and the parental cell line. Enrichment analysis revealed that associate gene ontology (GO) terms with these common DE genes are integral components of the plasma membrane and cytoskeleton (Benjamini-Hochberg adj. p-value < 0.05). Patterns of changes in the transcriptome data further confirmed observed differences in the genome analysis. They highlighted that the main differences between parental HEK293 clones and their progeny cell lines are related to cytoskeleton matrix organization and cell membrane structure and could be a selective phenotypic advantage for progeny cells during continuous cell lines cultivation (Figure 9A).

Due to the importance of the growth morphology of cell lines in industrial bioprocessing, I aimed to find differences between adherent and suspension cells that lead to the adaptation of adherent cells to grow in suspension cell culture. Pairwise differential expression analysis of clones resulted in the detection of 38 genes with a constant and significant (Benjamini-Hochberg adj. p-value < 0.05, absolute log2 fold change > 1) pattern of change between adherent and suspension cells (Figure 9B). Furthermore, enrichment analysis of these 38 genes suggested a potential difference in the cholesterol biosynthetic pathway between adherent and suspension cell types.

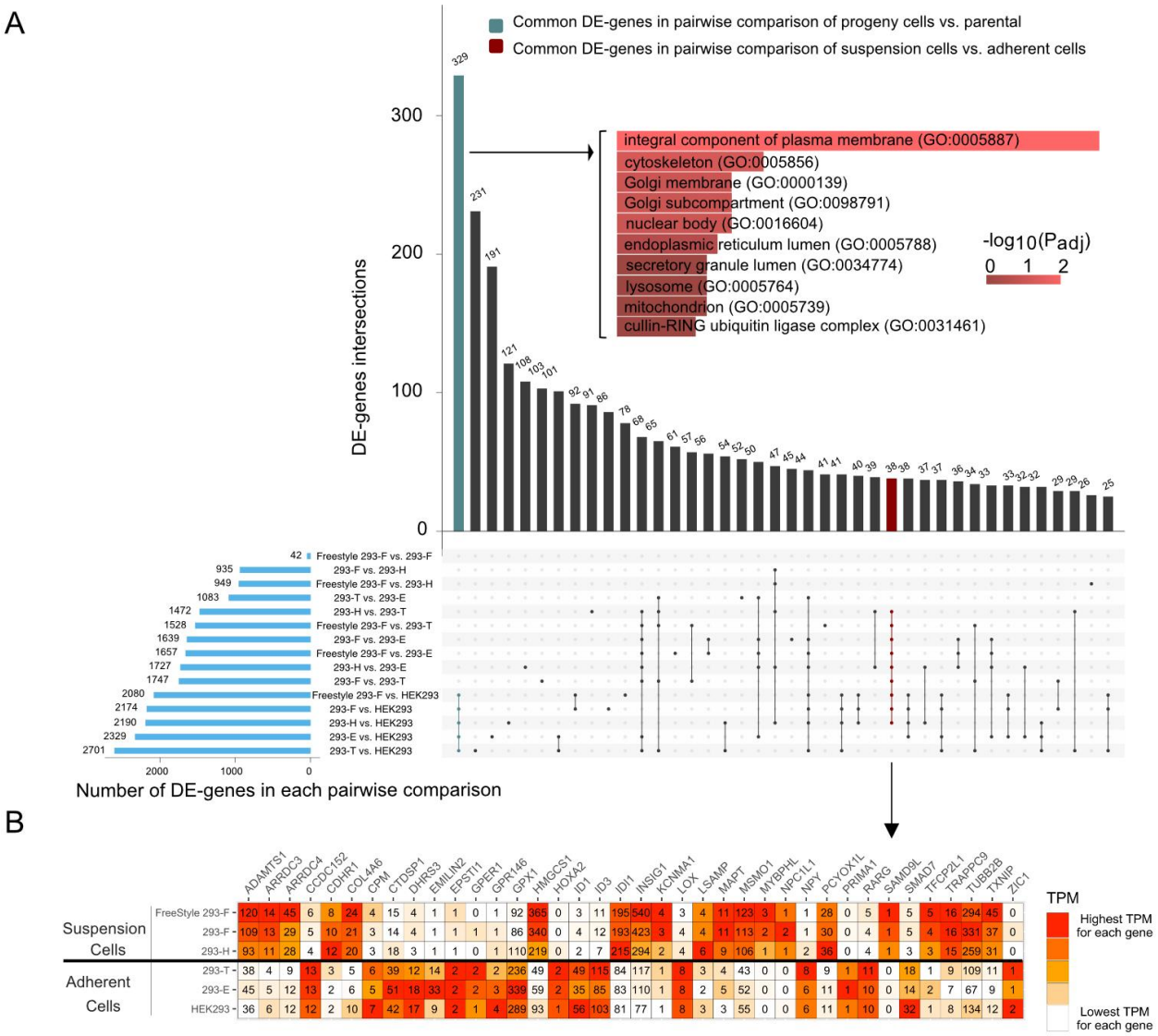

Figure 9. Pairwise comparison of transcriptomes of HEK293 cells. (A) In a comparison of progeny and parental HEK293 cells, 329 genes (green) were detected as commonly differentially expressed (DE) (B.H. adj. p-value < 0.05 and absolute log2 fold change > 1) among all comparisons. Enrichment analysis highlighted DE genes are associated with integral components of plasma membrane and cytoskeleton. (B) Thirty-eight DE genes (red in panel A) were detected as commonly differentially expressed in pairwise comparison of adherent and suspension cells and were enriched in those involved in the cholesterol pathway.

Focusing on the 38 DE genes between adherent and suspension cells, I aimed to find those genes that indicate a constant pattern of change between all adherent and suspension cells. For this purpose, I collected transcriptome data for 63 cell lines from the Human Protein Atlas (Thul et al. 2017) and grouped them based on their morphology, yielding 47 adherent and 16 suspension cells. Next, I sought to find the pattern of change for 38 DE genes detected in the previous step (Figure 10). Out of the 38 genes, five genes (LOX, ID1, ZIC1, DHRS3, and RARG) indicated a consistent and similar expression change pattern (all down-regulated) between adherent and suspension cells. These genes may play essential roles in the transition from adherent to suspension morphology. Results presented in this study could be beneficial for further cell-line engineering purposes and developing cell lines with a morphology switch to shift between adherent and suspension morphology in desired situations.

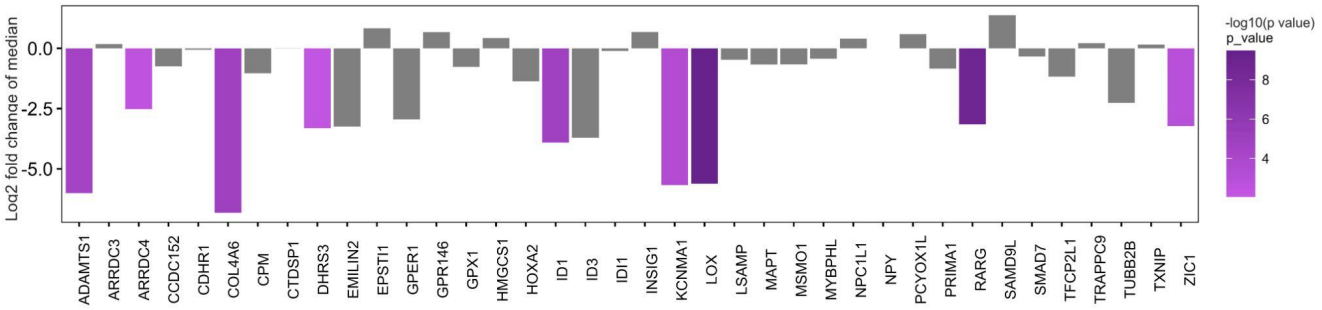

Figure 10. Evaluation of the differential expression pattern for 38 identified differentially expressed genes in 63 human cell lines. Five genes (LOX, ID1, ZIC1, DHRS3, and RARG) were detected with a consistent and similar pattern of change. Bars show the logarithmic fold change of genes between two groups of suspension and adherent cells, and color shows the level of significance. Gray columns represent genes with an insignificant (p-value > 0.05) change.

Cell demands for the production of secretory and non-secretory proteins are different (Paper II)

Higher efficiency of the production process and better quality of the products are ongoing cell factory engineering research (Tambuyzer et al. 2020). In addition, recent waves in the drug discovery field have led to the development of new drug proteins with more sophisticated structures and elevated levels of complexity in the production process (Kintzing, Filsinger Interrante, and Cochran 2016). However, there are still pending questions regarding the protein production and secretion process within the cells, such as: what are the differences in metabolic resource allocation for the production of secretory and non-secretory proteins (Gutierrez et al. 2020), how does energy metabolism need to be rewired to meet the energy requirements for the production of secretory and non-secretory proteins (M. Huang et al. 2017; Lodish et al. 2000), and how do cells allocate enzymatic resources to support the demand for production of proteins (Yu Chen and Nielsen 2019)? Comparing transcriptome profiles of cell lines that produce secretory and non-secretory proteins at different rates could help address the challenges in improving the quality and efficiency of the protein production process through designing improved cell factories. In this study, I established two groups of HEK293F cell lines, producing either secretory erythropoietin (EPO) or non-secretory green fluorescent protein (GFP) (Figure 11A-B).

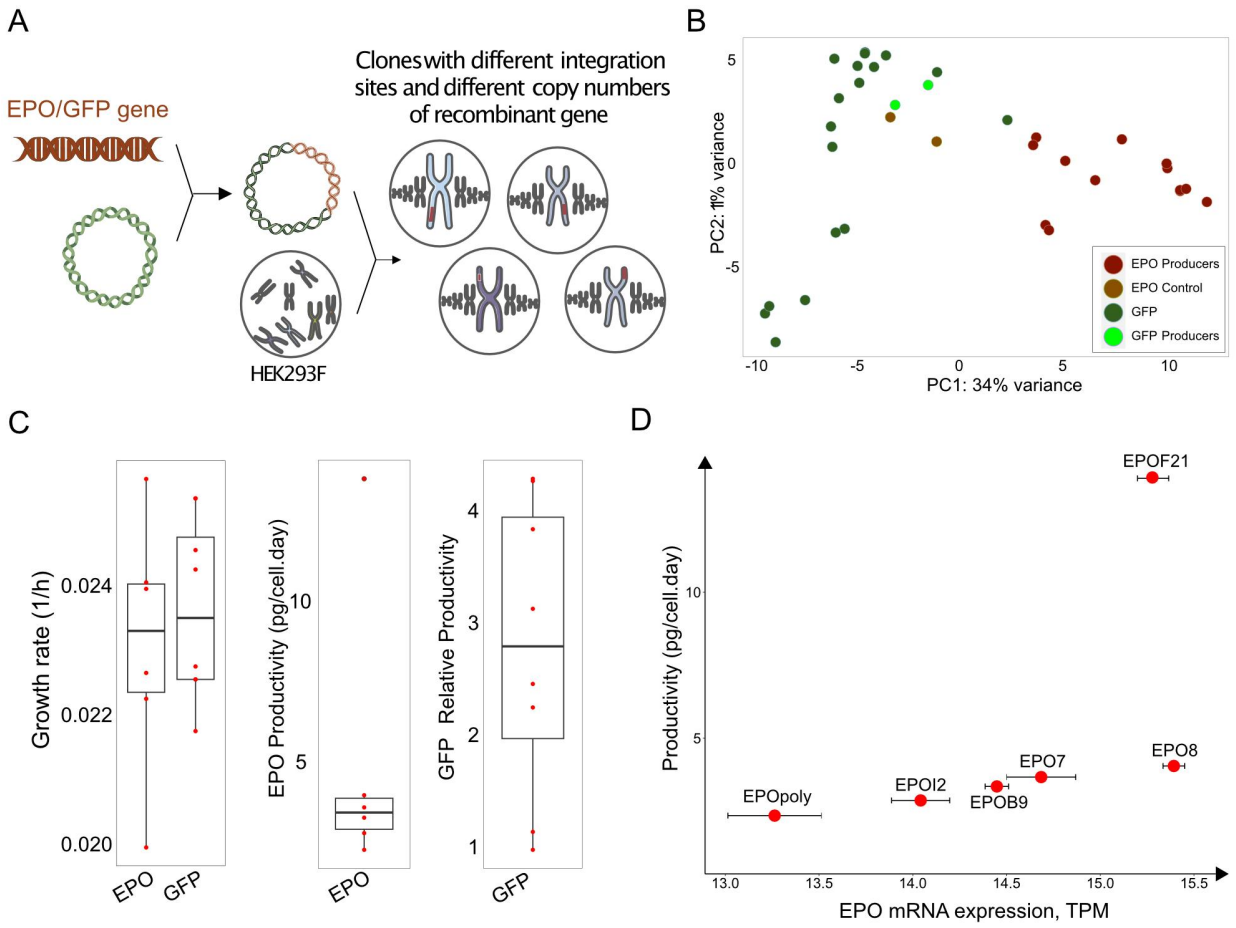

Figure 11. Cloning procedure for generating EPO and GFP producer clones. (A) The schematic diagram of the cloning procedure results in random integration of recombinant genes (EPO or GFP) and establishing clones carrying various copy numbers of recombinant genes and different protein production levels. (B) Transcriptome profiles separate EPO and GFP producer clones in the first component of principal component analysis (PCA). (C) The growth rate of clones producing EPO and GFP. Range of specific productivity between EPO producer clones and range of relative GFP production between GFP producers. (D) EPO mRNA expression versus specific productivity between EPO producer clones.

The growth rates of recombinant protein (r-protein) producer cell lines were lower compared to the growth rate of their parental cell lines, 22% and 14% in EPO and GFP producers, respectively. However, the protein production rate showed a more comprehensive range among clones; Five EPO producer clones exhibited almost a 6-fold difference in their specific productivity, whereas seven GFP producers showed up to a 4-fold change between the lowest and highest GFP producer (Figure 11C). Specific productivity of EPO production in EPOF21, the most productive EPO producer clone, was 13.9 pg/cell-day, which was over 3-fold higher than the second-highest EPO-producer clone EPO8 (productivity = 4.05 pg/cell.day, Figure 11D). Interestingly, comparing mRNA levels of the EPO gene between EPOF21 and EPO8 showed a 20% lower mRNA copy number in EPOF21 (Figure 11D).

Energy availability is necessary for recombinant protein production

By comparing the transcriptome profiles of cell lines producing secretory EPO protein with cell lines producing non-secretory GFP, I found 922 up- and 64 down-regulated genes (B.H. adj p-value < 0.05, $|Log2FC| > 1$). In addition, gene set enrichment analysis on the differential expression analysis results highlighted upregulation (B.H. adj. p-value < 0.05) of protein secretion-related pathways such as targeting proteins to the endoplasmic reticulum or cell membrane, as well as some other pathways such as oxidative phosphorylation (Figure 12A).

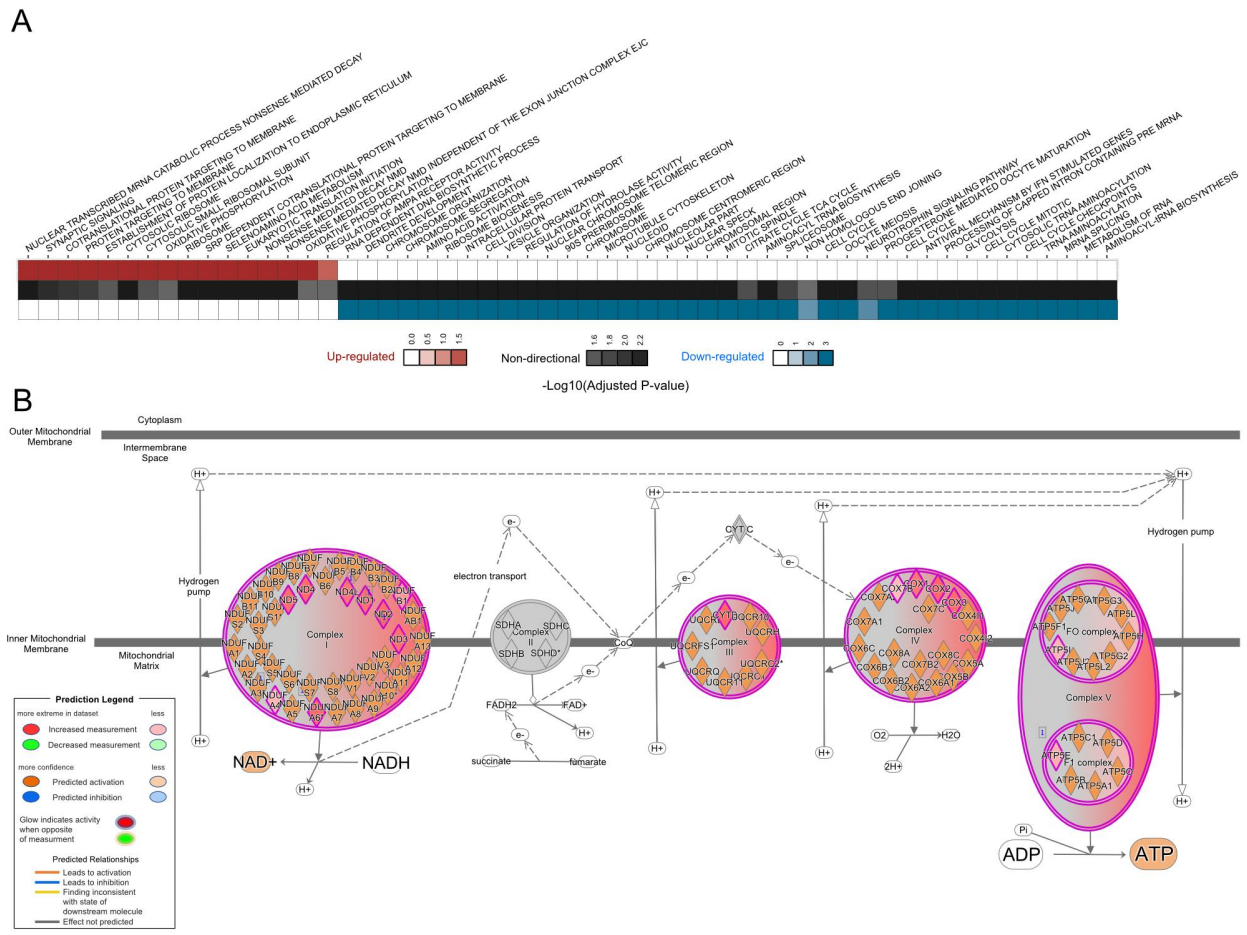

Figure 12. Production of secretory EPO needs higher levels of energy. (A) Pathways related to secretion and oxidative phosphorylation indicate upregulation in EPO producers. (B) Oxidative phosphorylation is upregulated in EPO producers cells, suggesting an increase in ATP production in these clones.

Further investigations of the oxidative phosphorylation pathway using Ingenuity Pathway Analysis (IPA) (Figure 12B) revealed that in EPO producers, all electron transport chain complexes except complex II were enriched up-regulated genes, which may support higher ATP production.

Ribosomal components adapt to meet protein production requirements

Another interesting observation in comparing EPO and GFP producers involved the pattern of expression change among ribosomal proteins. Ribosomal proteins were upregulated in both EPO and GFP producers (Figure 13A). Further investigations through pairwise comparison of each EPO and GFP clone with their corresponding controls (Figure 13A) indicated cell lines upregulate ribosomal proteins in a r-protein-specific pattern. I found common differentially expressed ribosomal genes (B.H. adj. p-value < 0.05, $|\text{Log}_2\text{FC}| > 0.58$) in a pairwise comparison of EPO and GFP producer clones with their corresponding parental cells (Figure 13A). Although I did not find proteins common in all pairwise comparisons of GFP producers with their parental control, EPO-producers showed 22 ribosomal genes that were always differentially expressed in EPO producers compared to their parental clone (Figure 13B). Functionality analysis of these genes showed that they are mostly related to SRP-dependent co-translational protein targeting the

ER, indicating a role of these genes in protein secretion. This observation suggests that ribosomal components are not fixed and confirms previous findings regarding the rearrangement of ribosomes under different situations to accommodate the translation of different mRNAs (Genuth and Barna 2018).

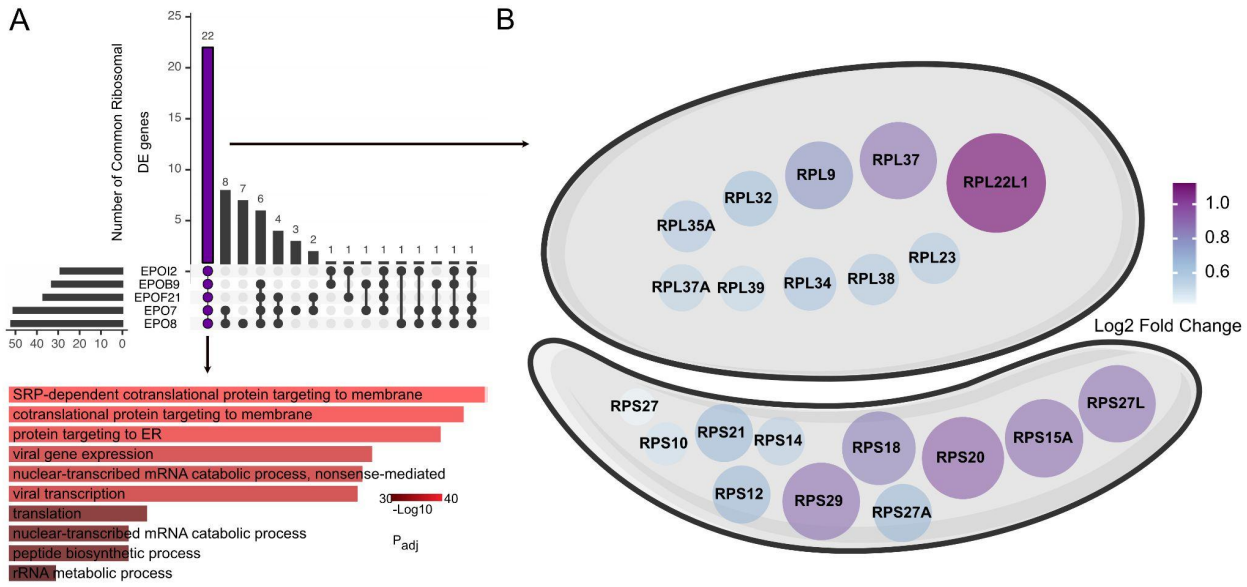

Figure 13. Ribosomal genes show a recombinant protein-specific pattern of upregulation in protein producer clones. (A) Number of common significantly differentially expressed (B.H. adj. p-value < 0.05, $|Log_2FC| > 0.58$) ribosomal genes between EPO-producers and the parental HEK293F cell. (B) Fold changes of differentially expressed ribosomal genes between EPO-producers and control.

The availability of multiple cell lines that produce either EPO or GFP but at different rates enabled us to investigate the correlation between the expression of each gene across clones with the expression of recombinant proteins in each group. I then investigated the pathways whose gene expression correlated with the level of recombinant protein production (Figure 14).

Mitochondrial ribosomal genes positively correlate with EPO secretion

By performing a correlation analysis I found 316 genes that showed either positive or negative correlation ($Mean_{TPM} > 10$, |Pearson's r| > 0.5, $p < 0.05$) with EPO production, and 118 genes correlating with GFP (Figure 14A). I found six genes that showed a significant correlation ($|r| > 0.5$, $p < 0.05$) with both EPO and GFP production, mostly involved in post-translational modifications and regulation of the protein production process (Figure 14B-C). GO term enrichment analysis (HyperGSA, $p < 0.05$) of genes correlated with EPO production indicated protein folding, post-translational protein modification, and post-Golgi mediated transport were among the pathways enriched with positively correlating genes (Figure 14D). Moreover, the mitochondrial organization was enriched with both positive and negative EPO correlating genes. Genes correlated with GFP production mostly showed a negative correlation pattern and were primarily associated with RNA catabolic processes and RNA export from the nucleus (Figure 14E). The protein translation pathway was enriched (HyperGSA p-value = 0.029) with positive and negative GFP-correlating genes.

Many of the genes that negatively correlated with both EPO and GFP production were ribosomal components. To further investigate, I investigated the correlation of all ribosomal proteins in both EPO and GFP-producing clones (Figure 14F-G). Interestingly, I found that cytosolic ribosomal genes (e.g., RPL38 and RPS9) exhibited a negative pattern of expression with increased EPO production. In contrast, mitochondrial ribosomal genes (such as MRS18A, MRPL49, MRPL40, and MRS11) showed a positive expression trend with increasing EPO production (Figure 14F). This is consistent with our previous observation of an upregulation of genes associated with the electron transport chain in EPO producers. Many upregulated genes in the electron transport chain have a mitochondrial origin, and mitochondria need ribosomes as translation machinery to express those genes. Thus, an increased expression of mitochondrial ribosomal genes favors higher energy production that could support higher EPO production. However, as a trade-off, it is associated with a downregulation of cytosolic ribosomal components.

I observed a significant shift in the energy metabolism of EPO producers compared to GFP producers and parental clones (Figure 12). Moreover, I identified a transgene-specific pattern of ribosomal protein expression (Figure 13), which could be a strategy that cells deploy to adapt to different situations, such as producing recombinant proteins. I also observed that cells deploy a strategy to allocate more protein resources to produce more mitochondrial ribosomes while decreasing the production of cytosolic ribosomes (Figure 14). Although cytosolic ribosomes are directly needed for r-protein translation, producing more mitochondrial ribosomes may lead to higher translation of electron transport chain proteins of mitochondrial origin. Consequently, higher ATP production could indirectly boost EPO production.

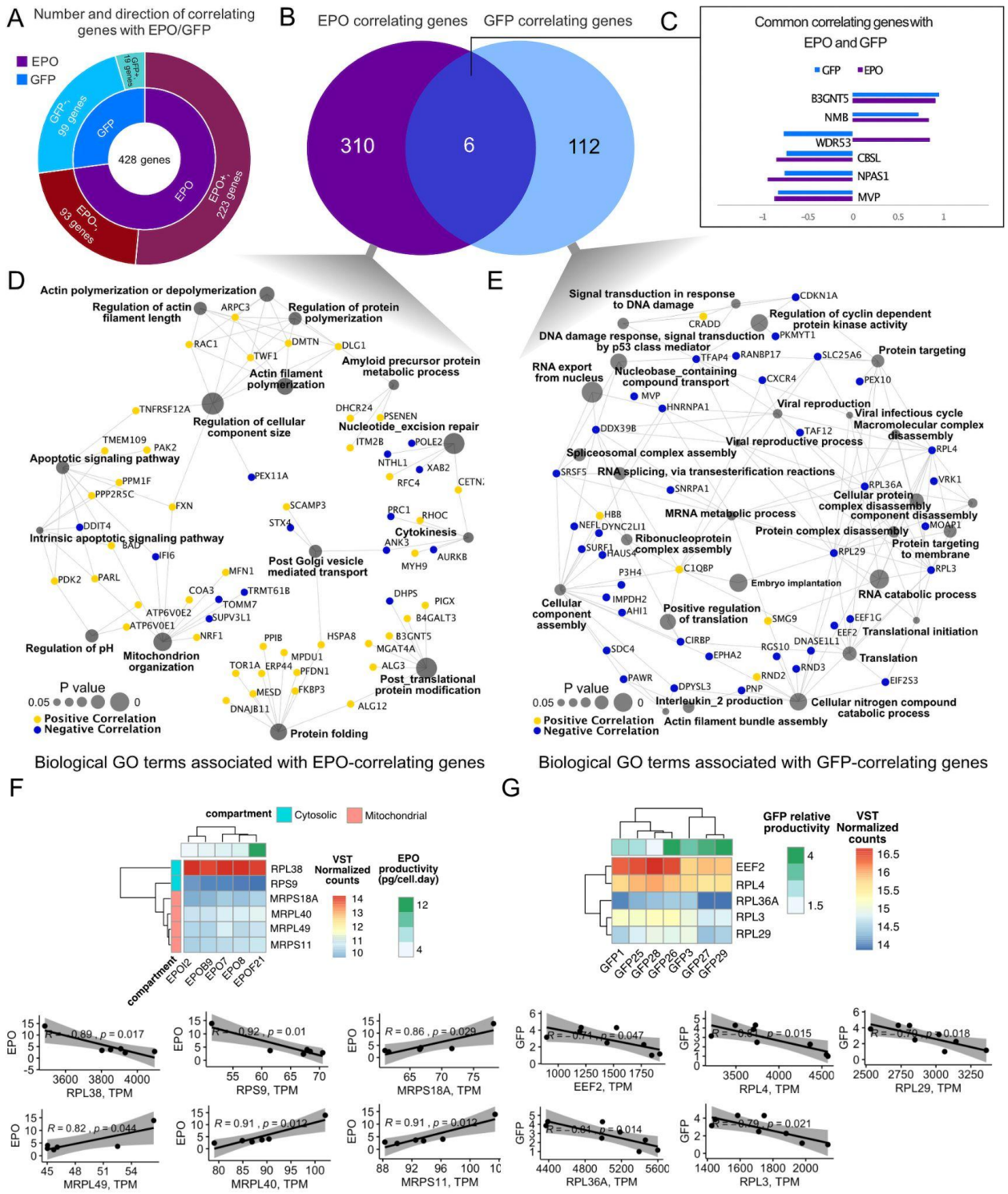

Figure 14. Correlation of gene expression with EPO and GFP production. (A) Negatively and positively correlating genes with EPO and GFP production. + means positive while – means negative (B) Common correlating genes with EPO and GFP. (C) Six genes exhibit a correlation between their expression and EPO or GFP production. (D) GO enrichment analysis of genes that are correlated with EPO production or (E) GFP production. (F) The trend of correlation for ribosomal genes whose expression correlates with EPO production or (G) GFP production.

Multi-omics of protein secretion by HEK293 cells (Paper III)

In the previous study, I compared each EPO producer clone with EPOF21 at the transcriptome level to capture unique patterns of changes that could potentially lead to higher EPO production in EPOF21. Transcriptome enrichment analysis confirmed that significantly changing pathways between EPOF21 and other lower EPO producers were related to post-translational pathways such as response to ER stress, response to topologically incorrect proteins and protein catabolic processes, as well as signaling and metabolic pathways such as mTOR signaling and carbohydrate metabolic processes (Figure 1G and Figure S6B in paper II). However, functional diversity of altered pathways suggested that the transcriptome cannot solely explain the mechanisms behind higher EPO production in EPOF21. Therefore, in a new project (paper III), I designed an experiment to comprehensively monitor molecular changes in different omics layers of EPOF21 (a high EPO producer clone), EPOI2 (a lower EPO producer clone), and HEK293F (the parental clone) (Figure 15A). I collected supernatant samples from each day of culturing for metabolomics analysis and samples from day four for transcriptomics and proteomics analysis (Figure 15A). I also measured the EPO production level for both EPO producer clones (Figure 15B). Both EPOF21 and EPOI2 hold two copies of the EPO gene but showed almost a two-fold difference in EPO production level (7.8 and 4.1 pg/cell.day, respectively, Figure 15B). This suggests potential differences in metabolic and signaling pathways between the two clones that confer higher EPO production to EPOF21.

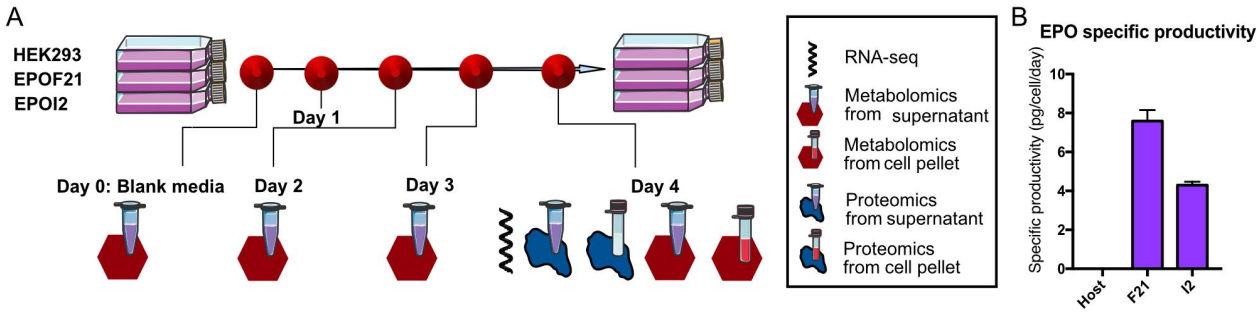

Figure 15. EPO production rate differs between EPO-producer clones. (A) Schematic of design of experiment (B) EPO production rates vary between EPOF21 (7.8 pg/cell.day) and EPOI2 (4.1 pg/cell.day).

Multi-omics analysis of cells producing EPO at different rates

In a pairwise comparison of transcriptome data between EPOF21 and EPOI2 (Figure 16A), TNFA signaling via NFKB, cholesterol homeostasis, unfolded protein response, PI3K-AKT mTOR signaling, mTORC1 signaling, p53 pathway, and UV response pathway showed significant (B.H. adj. p-value < 0.05) upregulation in the EPOF21 clone compared to EPOI2, while estrogen early response and apical surface and MYC targets were significantly downregulated (Figure 16B). Comparison of proteomics data between EPOF21 and EPOI2 (Figure 16C) followed by pathway enrichment analysis revealed upregulation of DNA strand elongation and replication processes (B.H. adj. p-value = 6.75e-06 and 9.52e-06 respectively) as well as protein processing in endoplasmic reticulum (adj. p-value = 3.42e-05) and protein export (adj. p-value = 5.88e-03) in EPOF21. Proteins associated with tRNA aminoacylation and aminoacyl-tRNA biosynthesis were expressed to a higher extent in EPOI2 (B.H. adj. p-value =

2.83e-06 and 3.38e-06, respectively). In contrast, proteins associated with the degradation of valine, leucine, and isoleucine showed lower expression in EPOF21 (adj. p-value = 1.49e-05) (Figure 16D). In the metabolomics data from day four of culturing (Figure 16E), metabolites involved in arginine and proline metabolism exhibited significantly different abundances in EPOF21 compared to both parental and EPOI2 clones (Figure 16F, $FDR_{F21/I2}$ = 3.94e-05 and $FDR_{F21/293F}$ = 8.42e-05). Furthermore, the same pattern of changes was observed for glycine and serine metabolism when comparing EPOF21 with EPOI2 ($FDR_{F21/I2}$ = 3.94e-05, Figure 16F).

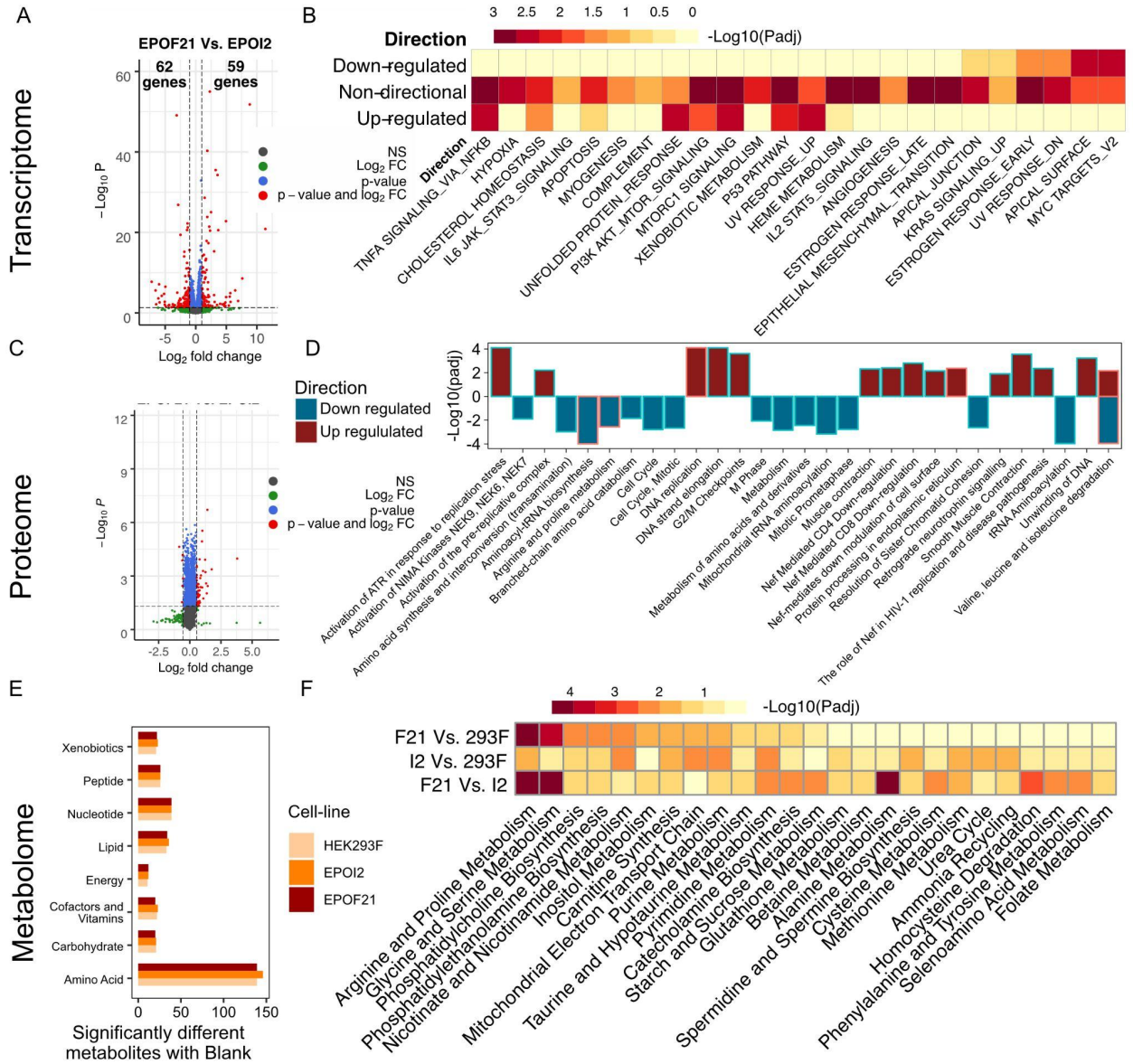

Figure 16. Comparative analysis of individual omics analysis. (A) Differentially expressed genes between high and low EPO producer clones. (B) Gene set enrichment analysis based on DE genes from the comparison of EPOF21 with EPOI2. (C) Differential expression analysis of EPOF21 and EPOI2 using proteomics datasets (D) Enriched KEGG and Reactome pathways based on differentially expressed proteins between EPOF21 and EPOI2. (E) The number of metabolites with significantly different abundance in the media on day 4 compared to blank media. (D) Enriched pathways with differentially abundant metabolites in pairwise comparisons of the three clones.

Investigating variables in principal component analysis

Clones separated by EPO production ability in principal component analysis (PCA) across all three blocks of omics data (transcriptome, proteome, and metabolome). In the first component of PCA analysis, EPOF21 and EPOI2 separated from the parental HEK293F clone, and the second principal component (PC2) separated EPO producers from each other (Figure 17A). To follow up, I further investigated variables in omics datasets that contribute most to the separation of clones (Figure 17B) in PCA. Top contributing variables in PC1 could highlight genes, proteins, and metabolites important for EPO production. The top contributing variables in PC2 potentially promote higher production of EPO in the EPOF21 clone.

CTSZ, a lysosomal cysteine proteinase active in terminal degradation of proteins in lysosomes (Aiba et al. 2018), showed the highest increase in EPOF21 in comparison to EPOI2 (B.H. adj. p-value < 1.55e-48, $Log_2FC_{F21/I2}$ = 8.82). Other upregulated genes in EPOF21 that showed substantial contribution in separation between EPOF21 and EPOI2 included ZN726, H2AJ, MYD88, and CYP7B1. These genes are involved in modulating gene expression through transcriptional regulation and protein folding (Craig-Mueller et al. 2020; Isermann, Mann, and Rübe 2020; Huntley et al. 2006).

In the proteomic data, eukaryotic initiation factor 4A-II (EIF4A2) showed the highest difference between EPO producer clones ($Log_2FC_{F21/I2}$ = 0.56, B.H padj = 0.03) and was the top contributing protein to PC2. Mitochondrial carbamoyl-phosphate synthase (CPS1) also exhibited significantly higher expression in EPOF21 ($Log_2FC_{F21/I2}$ = 0.52, B.H padj = 0.01). CPS1 catalyzes the rate-limiting reaction in the urea cycle, which acts in the production of glutamine, arginine, and creatine through the biosynthesis of carbamyl phosphate from carbonate and ammonia (Diez-Fernandez and Häberle 2017). Moreover, the positive effect of CPS1 upregulation on activating the mTORC1 signaling pathway has been frequently reported (Mossmann et al. 2019). It could be an indirect effect of higher arginine production that promotes the production of pyrimidines and creatine (Brosnan and Brosnan 2010; Shuyu Wang et al. 2015; Wyant et al. 2017). Among other top ten proteins that contributed to the second PC, Flotillin-1 (FLOT1) and general vesicular transport factor p115 (USO1) were upregulated in EPOF21 (FLOT1: $Log_2FC_{F21/I2}$ = 0.3 and USO1: $Log_2FC_{F21/I2}$ = 0.15, both B.H padj < 0.01). These proteins are involved in protein vesicle trafficking and protein translocation (Sohda et al. 1998). Ubiquitin-conjugating enzyme E2J1 (UBE2J1) was also upregulated in EPOF21 ($Log_2FC_{F21/I2}$ = 0.34, B.H padj = 0.01) and is involved in ER-associated degradation of misfolded proteins (Lenk et al. 2002).

In metabolomics, among the top ten metabolites contributing to the second PC, creatine showed a substantially higher abundance in EPOF21 ($FC_{F21/I2}$ = 26.25, B.H padj = 0.008). Moreover, cytidine, 2'-deoxycytidine, and 5-methylcytidine all exhibited significantly higher levels in EPOF21 compared to EPOI2 ($FC_{F21/I2}$ > 1.79 and B.H padj < 0.03), which suggests an activation of pyrimidine metabolism in EPOF21 relative to EPOI2.

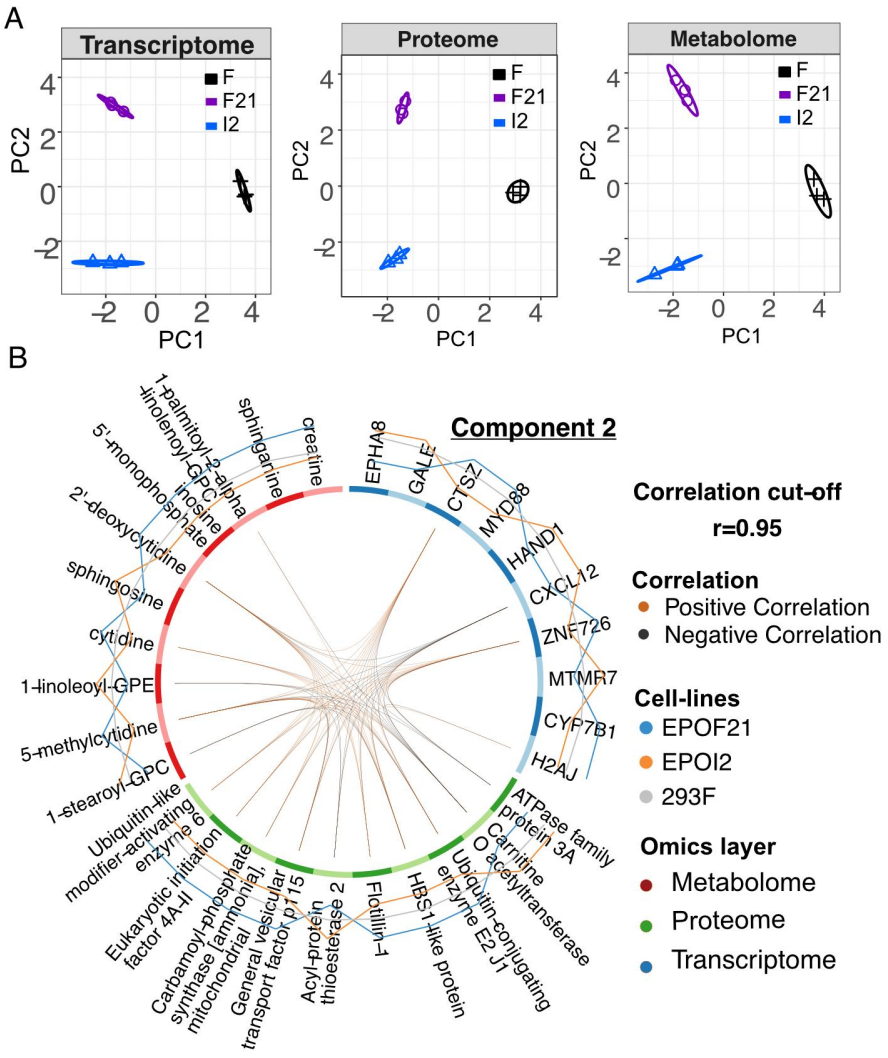

Figure 17. The principal component analysis separates cell lines based on EPO production. (A) EPO production ability separates cell lines in principal component 1 (PC1) in all three transcriptome, proteome, and metabolome datasets, whereas the EPO production level separates EPOF21 and EPOI2 in the second principal component (PC2). (B) Top ten contributing variables from each omics layer in PC2.

In comparisons of individual omics layers between EPOF21 and EPOI2, the transcriptome indicated activation of mTORC1 signaling in EPOF21. The metabolome highlighted the higher production of arginine-derived metabolites such as creatine and polyamines in EPOF21, and the proteome revealed higher protein processing in the endoplasmic reticulum in EPOF21 (Figure 18A-D). Likewise, further integrative analysis of omics data indicated top contributing variables in each omics layer are interconnected and suggested a general pattern of change in higher EPO producer clones. Among the top contributing proteins in PC2, I observed differences in CPS1 expression (Figure 18E), which catalyzes the first reaction in the urea cycle and could promote the production of arginine-oriented metabolites. Indeed, I observed higher levels of such metabolites (creatine and polyamines) in EPOF21 among the top contributing metabolites in PC2 (Figure 18A-B). These observations suggest activation of mTORC1 signaling pathways, which may occur due to the availability of arginine-derived metabolites such as polyamines. mTORC1 improves redox homeostasis within the cell and activates a group of downstream pathways, including protein production process and cell growth (Figure 18F).

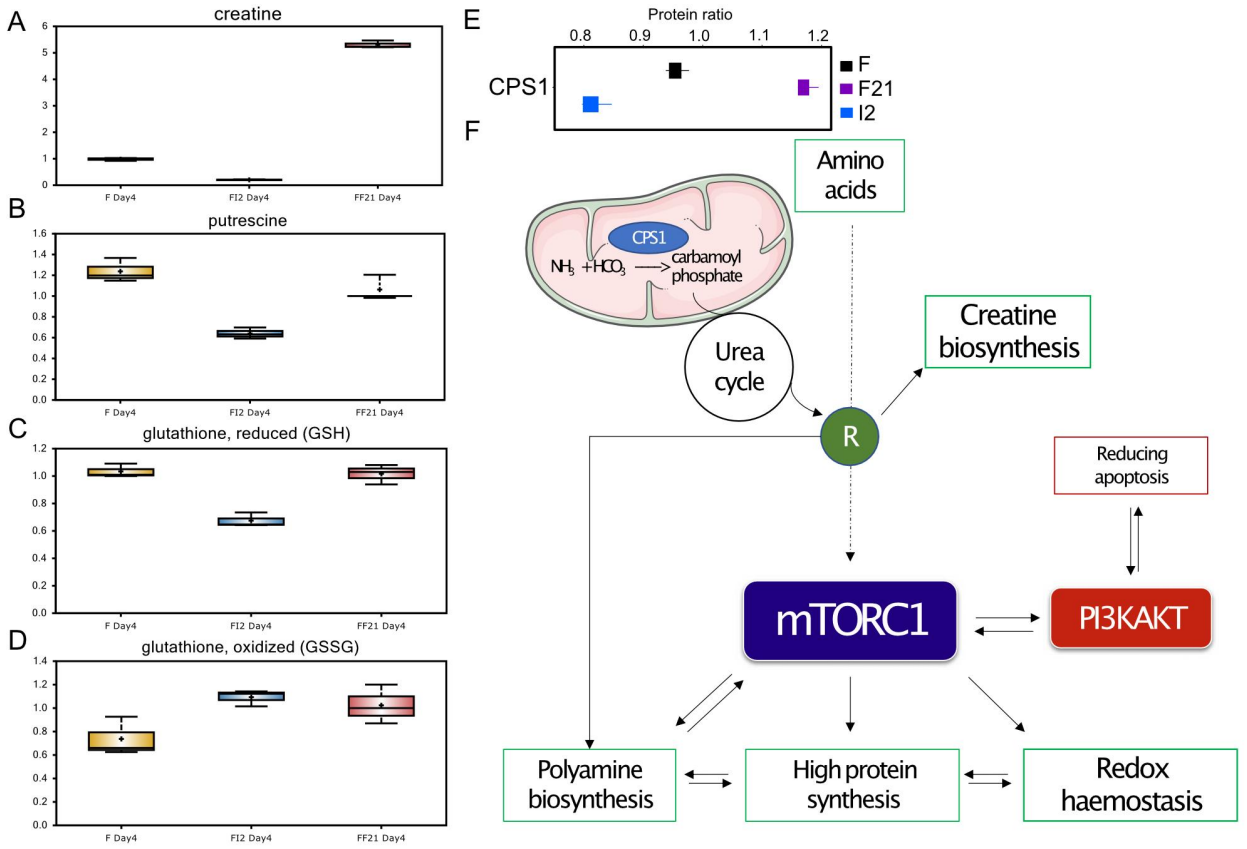

Figure 18. The general pattern of changes in protein and metabolite production activates the mTORC1 signaling pathway. (A-D) Arginine-derived metabolites such as creatine, polyamines (putrescine shown as one of the polyamines), and glutathione (E) CPS1, as the first enzymatic and rate-limiting reaction in the urea cycle have a higher expression in EPOF21 (F) suggested model of alterations in EPOF21 that lead to higher EPO production.

Predicting host cell proteins that compete with EPO secretion

Another approach to improve the production level of a recombinant protein (r-protein) in a cell line is to identify host cell proteins (HCP) that compete with the desired r-protein for metabolic, energetic, and enzymatic resources (Kol et al. 2020; Gutierrez et al. 2020). Finding HCPs that compete for most with the production of the r-protein could be performed using a constraint-based metabolic modeling approach. However, the most updated human reference genome-scale metabolic model (GEM), Human1 (Robinson et al. 2020), does not cover secretory pathways and cannot simulate the protein secretion processes. Hence, I drew on the methods from previous studies (Gutierrez et al. 2020; Feizi et al. 2013; Feizi, Banaei-Esfahani, and Nielsen 2015; Feizi et al. 2017; Robinson et al. 2019) and developed a systematic approach to add protein secretion reactions to the Human1 model.

Protein Secretion modeling by HumanSec toolbox

I first developed cell-line-specific GEMs using gene expression data (Mardinoglu et al. 2014). For generating models, I used the tINIT algorithm implemented in the RAVEN toolbox (H. Wang et al. 2018) and the Human1 model (Robinson et al. 2020) as the standard reference GEM for human cells (Figure 19A).

Protein specific information matrix (PSIM)

To generate protein secretion models, I used the approach developed by Feizi et al. (Feizi et al. 2013) for yeast and later used for CHO cells (Gutierrez et al. 2020). In this approach, I first collected characteristic information from UniProt for each protein and generated a table of retrieved information with proteins in rows and protein attributes as columns; this table was called the protein-specific information matrix (PSIM). The collected information includes UniProt ID, protein name, ER signal peptide, number of disulfide bonds, number of N-linked glycosylations, number of O-linked glycosylations, presence of a transmembrane domain, protein localization information, and protein sequence.

Generating a list of template reactions for the secretory pathway

A list of reactions was generated, covering all potential steps in the process of protein secretion. To generate such a list of reactions, I reviewed published information regarding components of the protein secretion pathways and their function in this process (Robinson et al. 2019; Gutierrez et al. 2020). As different proteins have different characteristics, such as PTMs and different cellular localizations, the series of reactions comprising the secretory pathway must be tailored to each specific protein. Therefore, the list of all potential reactions was named the template reactions list and acted as a library of secretory reactions that could provide the necessary reactions for the production and secretion of the protein depending on each specific protein.

Reconstruction of the secretory pathway in human cells

I developed the HumanSec algorithm to perform the following steps (Figure 19A): (I) Identify secretory proteins from proteomics data. The user can decide the criteria for considering a protein as secretory. In our analysis, I used the presence of an ER signal peptide as the definition of a secretory protein. (II) Generate a protein-specific list of reactions for each secretory protein using information contained within the PSIM and the list of template reactions. (III) Add the generated list of reactions for each secretory protein to a reference GEM. In our case, I used cell-line-specific GEMs generated by the tINIT algorithm. (IV) Finally, the algorithm performs a gap-filling step in case of a failure to produce any of the secretory proteins. In general, I added specific reactions for the biosynthesis and secretion of 559 proteins to GEMs (Figure 19B).

Out of 9,791 proteins, 6,497 were detected in cell pellet samples, and 2,861 proteins were detected in supernatant samples. The final location predicted for 168 proteins as extracellular, and 391 proteins were predicted to carry an ER signal peptide in their sequence, suggesting these proteins are ER and Golgi apparatus clients and need post-translational modification (Figure 19B). By generating protein secretion models that cover reactions for these proteins, 21,760 new reactions were added to each of the cell-line-specific GEMs.

All codes and functions for generating the PSIM matrix, template reactions list, reconstruction of the protein secretion model, and overlaying omics data on the protein secretion network map is archived in a toolbox called HumanSec that can be accessed from its GitHub repository: https://github.com/SysBioChalmers/HumanSec.

Constraint-based analysis of reconstructed secretory models

To guarantee the production of all secretory proteins by the protein secretion models generated for EPOF21 and EPOI2, I used the measured ratios between EPO and each secretory protein in the proteomics dataset as constraints by defining pseudo reactions connecting the production of each secretory protein to EPO production. By taking this approach and maximizing EPO production, the reconstructed protein secretion models produced EPO and all other secretory proteins in a ratio consistent with the protein abundance ratios measured in proteomics analysis.

I took a similar but more complex approach for constraining the models based on data acquired from metabolomics datasets. First, I merged the protein secretion models for EPOF21 and EPOI2 (Figure S8 in paper III) and generated one single model representing specific metabolite IDs for each cell line. Then, I grouped measured extracellular metabolites as either consumed or produced, based on their change trend in each day of culture. Finally, I defined a stoichiometric constraint for each extracellular metabolite that forced the two models to consume or secrete the metabolite at a ratio equal to that measured in the metabolomics data. Thus, this approach does not constrain the uptake or secretion flux of metabolites to a fixed value but instead constrains the ratio of these fluxes between the two cell types (for more details on this approach, please read method M9 in paper III).

Host cell proteins competing for most with EPO

After constraining the protein secretion models based on the proteomic and metabolomic data, I aimed to identify which secretory proteins exhibit the highest competition with EPO over cell resources. Knockout of such proteins (assuming that these proteins are not essential for fundamental cellular tasks) is expected to affect EPO production positively. For this purpose, I blocked the production of each secretory protein individually and maximized EPO production. The resulting list of proteins was sorted based on their effect on EPO production. Figure 19C shows the cumulative effect of the top 20 secretory proteins whose knockout may improve EPO production.

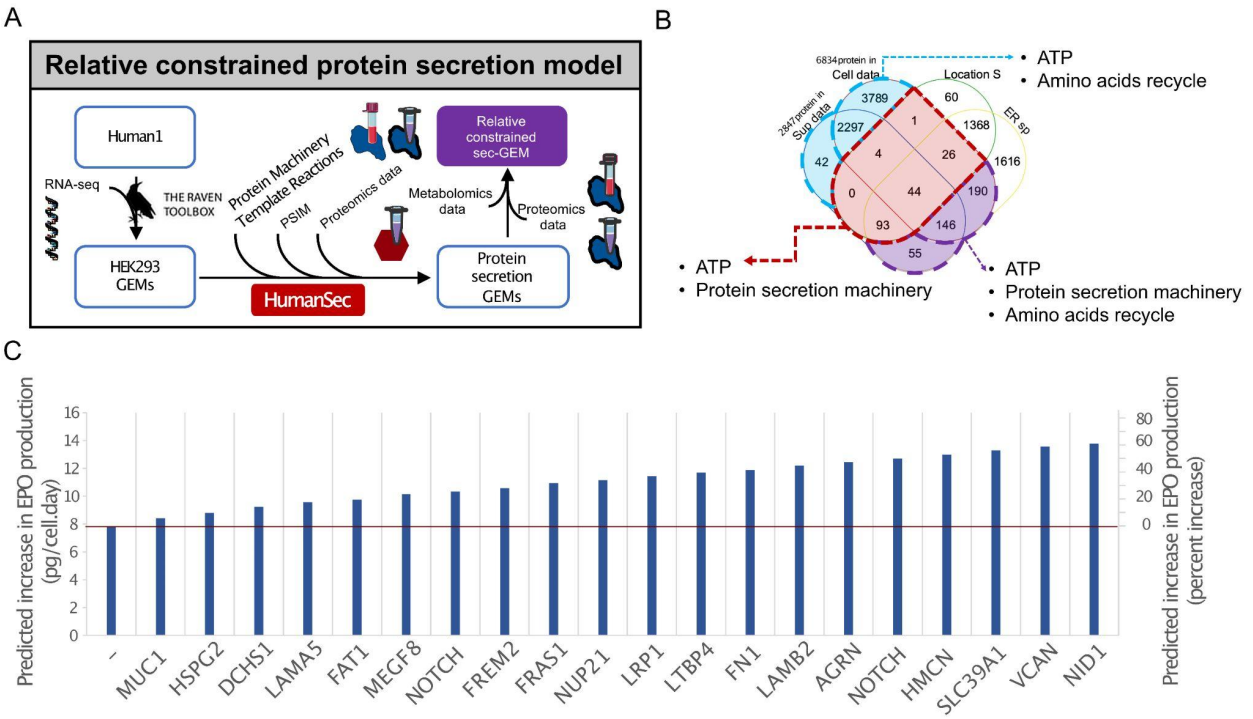

Figure 19. Genome-scale models assist with detecting host cell proteins that compete with EPO production. (A) Pipeline for generating protein secretion GEMs constrained by metabolomics and proteomics data (B) Categories of proteins detected in proteomics datasets. Proteins that were detected in our proteomics datasets and either carry an ER signal peptide (554 proteins) or have an extracellular predicted location (168 proteins) were included in the protein secretion model (559 protein in total, red and purple areas). (C) The predicted cumulative effect of knocking out the top 20 host cell proteins exhibiting the highest competition with EPO production for cellular resources.

This study identified the most significant alterations in different omics layers of a high EPO producer cell line compared to its parental clone and a low producer cell line. I suggested a hypothesis for connecting observed changes in the higher EPO producer clone after further testing and validation could be deployed for cell factory development purposes. Furthermore, I developed a constraint-based modeling approach and generated a metabolic model covering the protein secretion process. First, I constrained the protein secretion model using multi-omics data to improve model predictions. Next, I used protein secretion models for the higher EPO producing cell-line to predict host cell proteins whose knockout could improve EPO production. Finally, I provided the algorithm and resources for generating such models in a publicly available toolbox called HumanSec that could be used for similar studies to develop more efficient cell factories.

Analysis of the protein secretory pathway in cancer (Paper IV)

Alterations in the expression of protein secretory pathway (PSP) components may cause many abnormalities within the cell (Shiyu Wang and Kaufman 2012; Lebeaupin, Yong, and Kaufman 2020; Costa et al. 2020; Feizi, Banaei-Esfahani, and Nielsen 2015). It has been shown that the collection of proteins secreted from the cell (the secretome) has a great potential to be used as a reservoir for severe diseases biomarkers (Robinson et al. 2019; Welsh et al. 2003; Stastna and Van Eyk 2012), or even as a reservoir of targets for developing new drugs against such abnormalities (Ding et al. 2020; Khanabdali et al. 2016; Schaaij-Visser et al. 2013). However, all proteins in the secretome are produced and processed by a core set of protein components, called the protein secretory pathway (PSP). Feizi et al. previously demonstrated that PSP genes in healthy tissues are expressed in a tissue-specific pattern to meet the production demands of the secretome in each specific tissue (Feizi et al. 2017). Furthermore, previous studies included in this thesis confirmed the presence of particular patterns of PTMs and responses to unfolded proteins that exist in different cell lines (Saghaleyni et al. 2020).

Machine learning analysis of the cancer transcriptome

I retrieved the expression profile of 575 PSP genes (Feizi et al. 2017) from 11,053 RNA-seq profiles deposited in The Cancer Genome Atlas (TCGA) (Table 1). These samples were annotated with cancer type (33 different types) and the stage of cancer progression (from stage I to stage IV, Figure 20A). Eight different machine learning classifiers were trained and evaluated using a 10-fold cross-validation approach (Figure 20B) to (I) classify samples based on the expression profile of PSP genes and (II) score the level of importance of each feature (gene) in predicting sample classes (Figure 20C). I defined sample classes based on cancer status (normal or tumor) and the status of cancer progression from stage I to stage IV. A gene scoring approach was developed to predict the level of importance of each feature (gene) in predicting sample classes (Figure 20C). The machine learning algorithms used in this study were random forests (Ho 1995), highly randomized trees (ExTrees) (Geurts, Ernst, and Wehenkel 2006), adaptive boosting (AdaBoost) (Freund and Schapire 1997), extreme gradient boosted trees (XGBoost) (Freund and Schapire 1997), linear discriminant analysis (T. Chen and Guestrin 2016), lasso regression (Tibshirani 1996), ridge regression, and support vector machine (Boser, Guyon, and Vapnik 1992). These algorithms cover some of the most commonly used methods in machine learning studies on biological data (Tarca et al. 2007; Sommer and Gerlich 2013) and also span a good range of disciplines such as ensemble learning (random forests, ExTrees), regularized logistic regression (lasso and ridge regression), and boosting (AdaBoost, XGBoost). Furthermore, it is possible to calculate feature importance scores using all these algorithms.

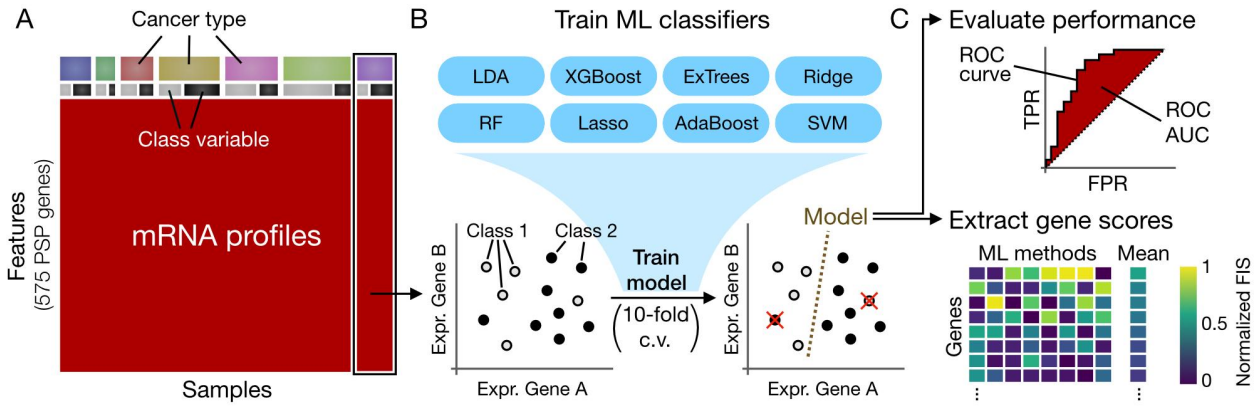

Figure 20. Schematic of machine learning analysis and feature scoring approach. (A) Expression data for 575 protein secretory pathway (PSP) components extracted from RNA-seq expression profiles of 11,053 samples from 33 different cancer types. (B) Eight different machine learning (ML) classifiers were trained using a 10-fold cross-validation approach to predict different binary classes, such as normal or tumor, and also different stages of the cancers, for example, stage I or stage II, etc. (C) ROC AUC curves used for evaluation of prediction performance for each model, as well as a feature importance score ranging from 0-1 calculated for each gene and ML algorithm and also averaged across all ML algorithms.

Machine learning approaches detect PSP genes that are regulated by P53

Out of the 575 PSP genes present in our analysis, four are directly regulated by the p53 protein encoded by the *TP53* gene and one of the most commonly mutated oncogenes in different cancer types (Ozaki and Nakagawara 2011). The four p53-regulated genes are *BCL2* Associated X (*BAX*), Heat Shock Protein Family A Member 4 Like (*HSPA4L*), Kinesin Family Member 23 (*KIF23*), and BCL2 Antagonist/Killer 1 (*BAK1*) (Graupner et al. 2011; Martin Fischer et al. 2013; M. Fischer 2017). Since a mutation in *TP53* is likely to affect the expression of these genes, by training ML models that classify *TP53* mutated and *TP53* non-mutated samples, I expected to find these four genes among the top important genes that contribute to the classification model. The average ROC AUC values across all ML algorithms classifying different types of cancer was 0.74 ± 0.11 (Figure 21A). The averaged consensus ML gene score across all cancer types identified three out of four genes that p53 directly regulates to have the highest consensus score among the 575 PSP genes (*BAX*, *HSPA4L*, and *KIF23*, Figure 21B).

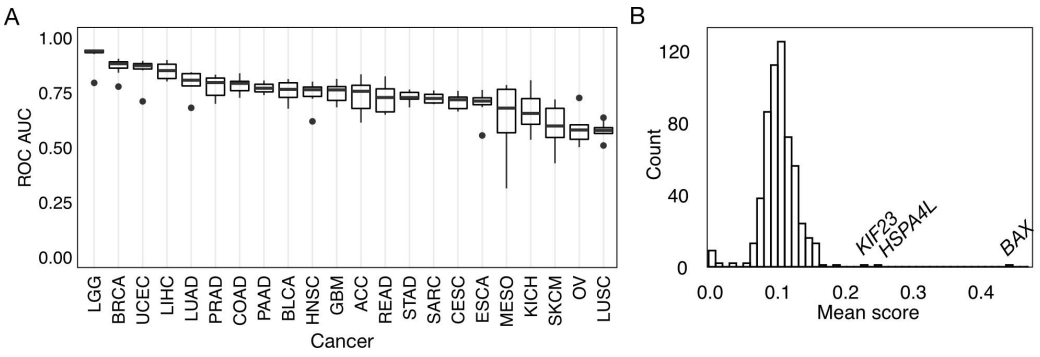

Figure 21. ML algorithms succeed in identifying genes that are regulated by p53. (A) Box plot of ROC AUC values categorized based on the cancer types. (B) Histogram of consensus gene scores calculated from ML analysis across all cancer types shows three out of the four PSP genes that are known to be regulated by p53 have the highest consensus score among all ML analyses.

Table 1. TCGA cancer abbreviations and metadata.

Code	Name	Cancer Status		TP53 Mutated		Tumor Stage			
		Solid Tissue Normal	Primary solid Tumor	False	True	Stage I	Stage II	Stage III	Stage IV
ACC	Adrenocortical Carcinoma	0	79	65	14	9	37	16	15
BLCA	Bladder Urothelial Carcinoma	19	408	214	194	2	134	147	142
BRCA	Breast Invasive Carcinoma	113	1090	637	336	201	689	274	22
CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	3	304	264	22	0	0	0	0
CHOL	Cholangiocarcinoma	9	36	33	3	26	10	1	8
COAD	Colon Adenocarcinoma	41	454	178	212	79	199	136	71
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	0	48	32	5	0	0	0	0
ESCA	Esophageal Carcinoma	11	161	29	131	21	71	51	8
GBM	Glioblastoma Multiforme	0	153	99	48	0	0	0	0
HNSC	Head and Neck Squamous Cell Carcinoma	44	500	154	340	27	86	86	278
KICH	Kidney Chromophobe	24	65	45	20	29	33	17	10
KIRC	Kidney Renal Clear Cell Carcinoma	72	530	323	9	291	68	139	102
KIRP	Kidney Renal Papillary Cell Carcinoma	32	288	273	5	187	23	63	19
LAML	Acute Myeloid Leukemia	0	0	0	0	0	0	0	0
LGG	Brain Lower Grade Glioma	0	510	273	229	0	0	0	0
LIHC	Liver Hepatocellular Carcinoma	50	371	254	105	190	98	97	6

Continuation of table 1. TCGA cancer abbreviations and metadata.									
Code	Name	Cancer Status		TP53 Mutated		Tumor Stage			
		Solid Tissue Normal	Primary solid Tumor	False	True	Stage I	Stage II	Stage III	Stage IV
LUSC	Lung Squamous Cell Carcinoma	49	501	89	400	270	179	89	8
MESO	Mesothelioma	0	86	68	13	10	16	44	16
OV	Ovarian Serous Cystadenocarcinoma	0	374	25	247	0	0	0	0
PAAD	Pancreatic Adenocarcinoma	4	177	63	107	21	150	3	5
PCPG	Pheochromocytoma and Paraganglioma	3	178	178	0	0	0	0	0
PRAD	Prostate Adenocarcinoma	52	495	434	56	0	0	0	0
READ	Rectum Adenocarcinoma	10	165	30	102	34	53	53	26
SARC	Sarcoma	2	259	145	90	0	0	0	0
SKCM	Skin Cutaneous Melanoma	1	103	398	67	77	140	171	24
STAD	Stomach Adenocarcinoma	32	375	197	175	59	126	156	42
TGCT	Testicular Germ Cell Tumors	0	134	127	1	56	12	14	0
THCA	Thyroid Carcinoma	58	502	485	2	321	59	125	61
THYM	Thymoma	2	119	114	4	0	0	0	0
UCEC	Uterine Corpus Endometrial Carcinoma	23	543	329	196	0	0	0	0
UCS	Uterine Carcinosarcoma	0	56	4	52	0	0	0	0
UVM	Uveal Melanoma	0	80	80	0	0	39	36	4

PSP genes associated with malignant transformation

The high scores of the three p53-regulated genes when classifying p53 mutated and p53 non-mutated samples confirmed that the ML approaches could capture relevant biological features associated with a phenotype. I next sought to train models for classifying normal vs. tumor samples and find the most relevant genes contributing to this classification (Figure 22A). By comparing the results of finding the most important genes detected through machine learning (ML) analysis with the results of differential expression (DE) analysis (Figure 22B), I found similarities and differences between the two approaches. The ROC AUC values indicating predicting the performance of normal vs. tumor based on the expression of PSP genes were substantially higher (0.97 ± 0.03) than the analogous analysis for detecting p53 mutated vs. p53 non-mutated samples (Figure 2 of paper IV). The higher prediction performance of ML methods in comparing cancer vs. normal compared to p53 binary mutation analysis is likely due to the greater physiological differences between normal and cancer cells than between tumor cells differing in a single gene mutation.

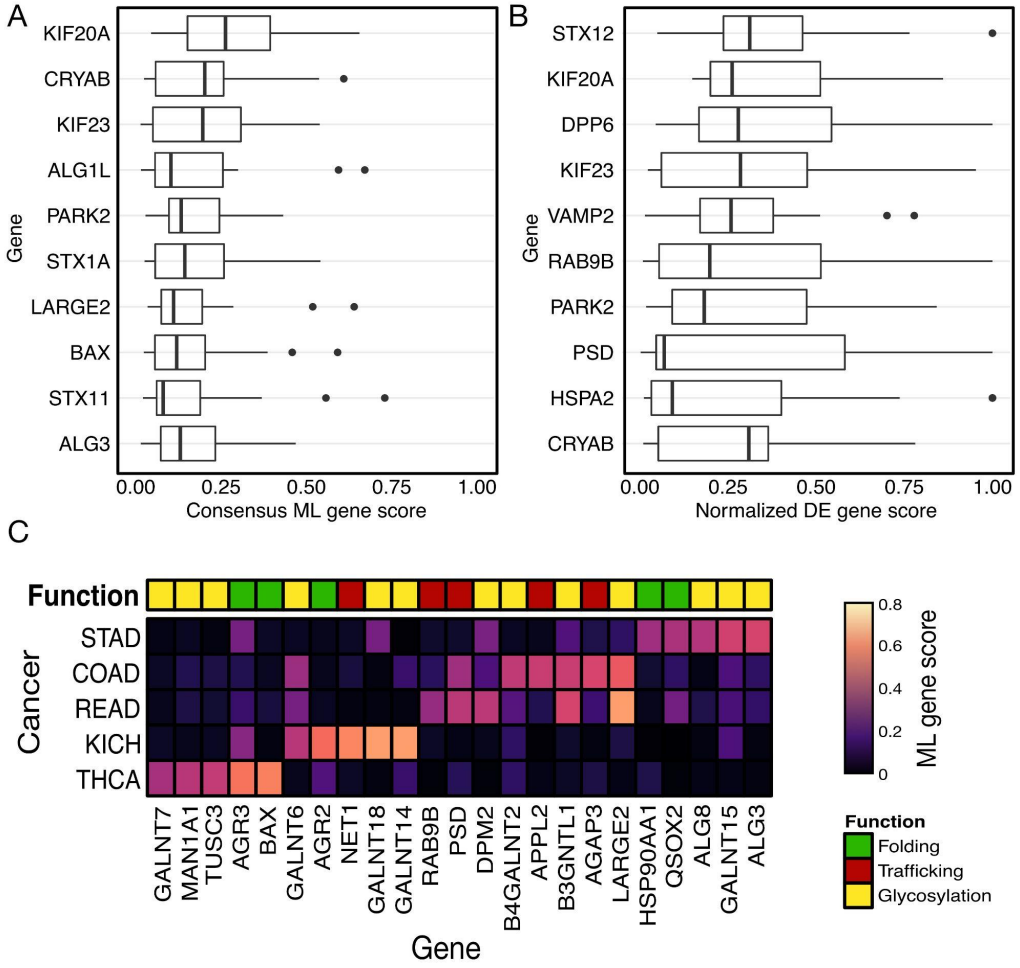

Figure 22. Glycosylation is enriched among top-scoring PSP genes from ML analysis for a subset of cancer types. Top ten scoring genes from (A) ML analysis of normal vs. tumor samples and (B) differential expression analysis of tumor against normal samples. (C) Glycosylation activity is enriched among the top PSP genes from ML analysis (3 out of 5 top genes in 5 different cancer types: STAD, COAD, READ, KICH, and THCA).

Investigation of the top-scoring genes by the ML and DE approaches revealed kinesin-6 family proteins (*KIF20A* and *KIF23*), Crystallin Alpha B (*CRYAB*), and several genes belonging to the soluble N-ethylmaleimide-sensitive-factor attachment protein receptor (SNARE) family (*STX1A*, *STX12*, *STX11*, and *VAMP2*) generally scored highly in both ML and DE approaches among the different cancer types (Figure 22A-B). *KIF20A* and *KIF23*, two top-scoring genes in the ML analysis (Figure 22A) that also exhibited significant (FDR adjusted p-value < 0.01) expression increase in comparing tumor and normal (expression increase in 16 out of 18 cancers, Figure 22A), are already under investigation in clinical trial studies, as targets that whose inhibition may help in stopping cancer progression (Rath and Kozielski 2012). *KIF20A* and *KIF23* are involved in Golgi-to-ER retrograde transport and have critical regulatory roles in mitosis and cytokinesis (Baron and Barr 2015; Lai et al. 2000). These results indicated the ability of ML methods in detecting genes that stood out in comparative DE analysis and suggested that other genes that were only detected by ML approaches could potentially have critical roles in cancer progression. For instance, from the STX proteins family, I detected *STX12* in the results of DE analysis, but *STX1A* and *STX11* scored between top genes in ML analysis (Figure 22A). The STX family supports various potentially tumorigenic functions such as autophagy and cell division (Jahn and Scheller 2006). Hence, these proteins could serve as targets with potential for further anti-cancer therapy studies (J. Meng and Wang 2015).

Besides several genes (*STX1A*, *STX11*, *CRYAB*) that were among top-scoring genes by averaging over all cancer vs. normal comparisons, other top-scoring genes did not exhibit consistent high scores in more than a few of the cancer types. For instance, almost all ML algorithms predicted a high score (consensus score = 0.82) for RAS oncogene family member 17 (RAB17) in analyzing prostate adenocarcinoma (PRAD) cancer samples. The RAB proteins family regulate vesicle trafficking and promote and suppress tumor growth, depending on the family member and cancer type (Gopal Krishnan et al. 2020).

Investigating the highest-scoring genes across all cancer types revealed a high frequency of genes associated with glycosylation. In particular, in five cancer types, STAD, READ, COAD, KICH, and THCA, 3 out of 5 top-scoring genes are known to participate in glycosylation activity (Figure 22C). This highlights the potential role of a protein's glycan pattern in dictating cellular functions such as cell adhesion, differentiation, and signal transduction that could promote cancer progression (Cummings and Pierce 2014).

PSP genes associated with tumor stages

I grouped tumor samples into four stages: I, II, III, and IV. To avoid the oversimplification and to assume an equal, linear change in disease severity of the different tumor stages, I took a binary approach to perform pairwise comparisons of different stages (e.g., I vs. II, I vs. III, II vs. IV, etc.) within each cancer type by training ML classifiers analogous to the same binary classification scheme that I used in the previous sections of this study. I performed the analysis on 20 cancer types, excluding cancer types without at least ten samples in at least two different tumor stages.

As expected, the class prediction performance (ROC AUC) for tumor stage predictions was substantially lower than those trained to predict normal vs. tumor samples (Figure 23A). The difference in the physiological status between different stages of cancer is relatively subtle compared to differences between normal and tumor samples. However, the distribution of ROC AUC values showed a cancer-specific pattern. Therefore, I only selected the trained models for the ten cancer types with the highest ROC AUC values for further investigation.

Across all ten different cancer types, *KIF20A,* on average, indicated the highest score (Figure 23B), particularly for separating tumor stages of KIRP, and to a lesser extent, KIRC and ACC. Although *KIF20A* in most cancer types showed either a higher ML score or a significant change in DE analysis in renal carcinomas (KIRP, KICH, and KIRC), I did not observe such a pattern. To investigate this behavior, I compared the expression of *KIF20A* across different tumor stages and cancer types, along with normal samples (Figure 23C). In most cancer types, *KIF20A* indicates a more step-like pattern of change between normal and cancerous cells, while in renal carcinoma samples, *KIF20A* expression increases gradually with increasing tumor stage. *KIF20A* has been highlighted in other studies as one of the risk factors involved in developing many different cancer types (Rath and Kozielski 2012; W. Zhang et al. 2016; Gasnereau et al. 2012). However, its involvement in renal carcinoma has not yet been addressed. The *KIF20A* expression alterations observed in the current study suggest that this gene may support more invasive and metastatic functions associated with later stages of renal carcinoma and thus constitutes a potential therapeutic target for this cancer type.

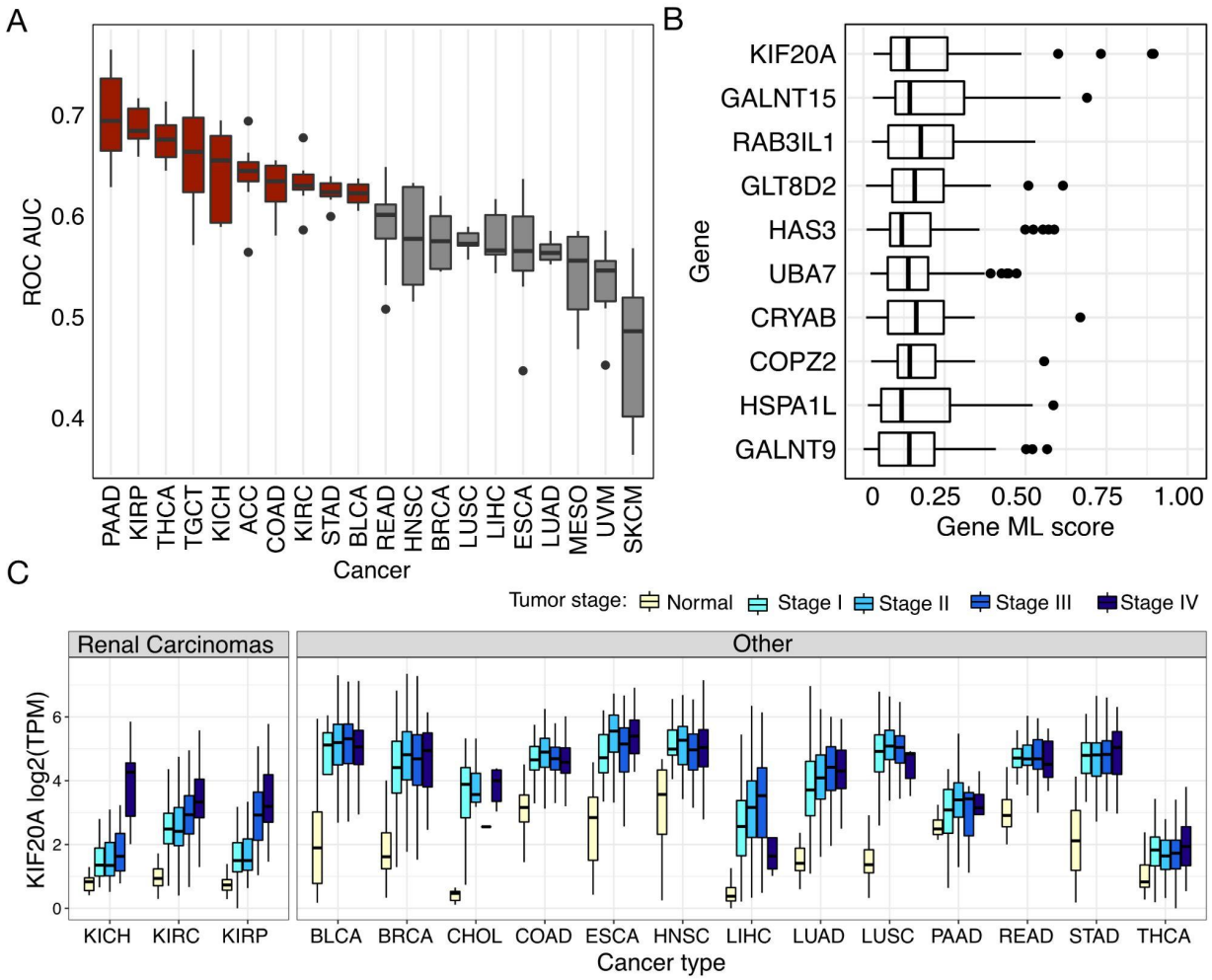

Figure 23. PSP genes associated with tumor stage. (A) Distribution of ROC AUC values for the prediction of tumor stages in each cancer type using PSP gens. (B) Top-scoring genes across all ML algorithms and cancer types. (C) *KIF20A* expression (log2 transformed TPM) among different cancer types and grouped by tumor stage.

In this study, I implemented a dual ML–DE approach to analyze 575 PSP genes in 11,053 healthy and cancerous RNA-seq profiles spanning 33 different cancer types. Besides investigating the selective potential of PSP genes in cancer diagnosis, selecting this subset of genes had the benefit of reducing the number of features for ML analysis. I validated the performance of our developed ML pipeline by detecting known targets of P53 among the 575 PSP genes (*BAX*, *HSPA4L*, and *KIF23*).

I then used the pipeline to find critical genes that had the highest contribution in classifying normal and cancerous samples. Through this analysis, I identified patterns in PSP expression that were common to carcinogenesis independent of cancer type. However, I also explored secretory elements that indicated a cancer-type-specific behavior. Finally, by training models for classifying samples from different tumor stages for each specific cancer type, I assessed the potential of ML approaches to identify genes relevant to tumor progression and potential for cancer treatment applications.

A recurring gene in our analysis was *KIF20A*. This gene was detected as top-scoring in nearly all ML and DE analyses: P53 mutated vs. P53 non-mutated, tumor vs. cancer, and comparison of

tumor stages. This observation is supported by the many different studies on the potential roles of *KIF20A* in cancer progression and tumor aggressiveness (W. Zhang et al. 2016). In addition, I detected a different pattern of expression for the *KIF20A* gene in samples from renal carcinomas (KICH, KIRP, KIRC). *KIF20A* exhibited a sharp difference in expression between normal and cancerous samples in most cancer types and remained relatively constant among different tumor stages. However, in renal carcinoma samples, I observed a gradual increase in the expression of this gene with increasing tumor stage. These observations support the further exploration of *KIF20A*-based anti-cancer treatments, though the activity of such treatments may differ in renal carcinomas.

Future Perspectives

The behavior of cells, as the smallest functional and structural unit of living systems (Jensen-Jarolim 2014), depends on the status of their internal components and the surrounding environment (Berg, Tymoczko, and Stryer 2002). Hence, predicting cell behavior requires accurate modeling of its inner components and must consider the effect of environmental conditions (B. Palsson 2006). However, there is an enormous number of reactions and interactions taking place within a cell. Genome-scale metabolic models (GEMs) greatly help with organizing the information on metabolic reactions within the cell in a unified manner (B. Palsson 2006), and flux balance analysis (FBA) approaches enables predictions of metabolites concentration changes and reaction fluxes (Orth, Thiele, and Palsson 2010). However, GEM reconstruction and FBA simulations involve many assumptions that are not always explicitly stated (O'Brien, Monk, and Palsson 2015). Although these assumptions enable computational simulations, they simplify the metabolic and physiological characteristics of the cell (Fang, Lloyd, and Palsson 2020). For example, many pathways such as transcription, translation, signaling pathways, protein secretion, and apoptosis are usually not included in classical GEMs. Furthermore, classical GEMs do not consider some of the spatiotemporal and physicochemical limitations; For instance, limitations in protein allocation for each reaction or limitations could affect reaction fluxes due to resource or spatial constraints.

Omics approaches can greatly improve our knowledge about the different components within cells from a global perspective. For instance, performing the transcriptomic analysis can provide information about all RNA transcripts in a cell while also providing comparative information about the relative levels of each transcript (B. Wang et al. 2019). The same applies to proteomics and metabolomics analysis (Di Girolamo et al. 2013). This information can help address the challenges of generating more advanced GEMs, with fewer simplifications and more accurate biological predictions (Hyduke, Lewis, and Palsson 2013). However, in most cases, information retrieved from omics analysis cannot be directly incorporated into GEMs (Yizhak et al. 2010). Thus, contextualizing omics information using GEMs needs proper approaches that can either use this data for increasing the coverage of a GEM's reactions or for extracting information that could be used as biological constraints (Yizhak et al. 2010; Dahal et al. 2020; Volkova et al. 2020; Ramon, Gollub, and Stelling 2018). The knowledge provided by omics datasets has been used so far in different studies to improve GEM coverage and predictions (H. Wang et al. 2018; Domenzain et al. 2021; Pandey, Hadadi, and Hatzimanikatis 2019). However, there is still a lack of unified methodology for reconstructing GEMs using multiple different omics datasets. Ideally, datasets such as transcriptomics, proteomics, and metabolomics could be used as input to generate a GEM with high reaction coverage and biologically accurate flux constraints.

One major challenge in developing such an algorithm is that the data in some omics approaches such as metabolomics and proteomics are reported as relative values between samples, which are not compatible with most current standard methods for generating GEMs constraining reactions performing FBA. Hence, to integrate data from comparative omics datasets into GEMs, we need to develop improved approaches for estimating the absolute values from relative values (Sánchez et al. 2021) or by improving algorithms that can reconstruct GEMs and incorporate FBA

constraints in a comparative mode for two or more samples. In the current thesis, I deployed the latter approach to predict the relative production of secretory proteins between two models (paper III). The promising results of this analysis could be a positive sign for developing an approach that could take into account the corresponding values for all proteins and metabolites and constrain similar reactions in two models based on these data. However, applying this approach on enzymatic proteins with complex promiscuous or isozyme relationships will further complicate the approach. Although expanding this approach to cover enzymatic reactions demands further formulation, it could be a way of reconstructing more enriched GEMs with more precise flux distribution predictions.

In a part of the current thesis, I described a pipeline for generating human GEMs that cover reactions in the protein secretory pathway. However, this pipeline is not able to predict the pattern of PTMs on each protein. As PTMs have a critical role in protein functionality, a possible way of expanding protein secretion models could be to expand them to cover and predict PTM patterns. For this purpose, we need to keep in mind that environmental conditions such as culture media and culturing processes and the status of signaling pathways within the cell could substantially affect the variability of alterations in the patterns of PTMs, presenting a challenge in modeling this process.

Tegel et al. indicated (Tegel et al. 2020) CHO and HEK293 cells have different power for producing various human secretory proteins. Thus, another potential application of protein secretion models might be using cell-factory-specific protein secretion models to find the most efficient cell factories to produce each protein of interest. Furthermore, by reconstructing such models and applying proper constraints based on culture media composition and culturing conditions, we could capture and learn how each cell factory's metabolism supports the production of different proteins. Solutions provided by protein secretion GEMs could greatly optimize the variables in the culturing process, such as culture media composition, and also be used for finding targets for cell factory design.

A well-constraint protein secretion GEM could be a powerful tool for analyzing samples from diseases that the protein secretion process plays a vital role in, e.g., Alzheimer's diseases and Parkinson's diseases (L. C. Walker and LeVine 2012). Although there are different hypotheses regarding the leading cause of deficiency in protein production and secretion in such disorders, the main reason is unknown (Long and Holtzman 2019). However, various studies have shown metabolism-related deficiencies, such as shortage in the power of energy production and imbalance redox condition, could have a significant role in causing such diseases (Long and Holtzman 2019; W. T. Wang et al. 2019; Ferreira et al. 2010). Protein secretion GEMs connect metabolism to the protein secretion pathway. So, these models potentially could serve as powerful tools for analyzing proteopathy disorders and finding biomarkers and drug targets in the metabolism that lead to the progression of proteopathy disorders.

Conclusion

Protein secretion is a multi-compartmental process catalyzed by many reactions within the cell and changes under various situations such as the production of a recombinant protein or disease. In the current thesis, I studied the protein secretion pathway in human cells using systems biology tools.

In paper I, I compared multiple strains of HEK293 cells, a recombinant protein production cell factory, to discover the key differences that could lead to better adaptation of cell lines from adherent to suspension culturing. I identified genes with a high potential to regulate this process.

In paper II, I compared cell lines that were producing a recombinant secretory protein with those that were producing a non-secretory protein. I observed that alterations in energy metabolism appeared to support higher protein production. I showed that higher protein producer cells dedicate more protein resources to overexpress mitochondrial genes that are involved in the electron transport chain, and through this provide more energy resources for recombinant protein production.

In paper III, I compared transcriptome, proteome, and metabolome between EPO-producer clones that produced EPO at different rates and investigated genes, proteins, and metabolites that potentially play more critical roles in the protein production process. These analyses suggested an important effect of activating the mTORC1 signaling pathway to boost protein production and reducing apoptosis. I expanded the human GEM to cover the protein secretion process in human cells and predicted proteins whose synthesis competes for metabolic resources with our recombinant protein of interest. To integrate relative metabolite abundance data for boundary metabolite constraints, I developed an approach that constrains the models based on relative protein abundances. I used a similar approach for constraining the production of secretory proteins by the model. These efforts could be continued in future studies for defining an approach that can implement relative values from proteomics and metabolomics analyses as soft constraints on the model's reactions and improve predictions.

In paper IV, I investigated the expression profiles of protein secretion pathway genes across samples from 33 different cancer types. I used a machine learning approach to predict important genes associated with differences between tumor and normal samples, as well as between different tumors. Our analyses in this study highlight: (I) PSP genes could potentially serve as a source for new anti-cancer drug targets, and (II) in analyzing RNA-seq data, applying different analytical approaches can supplement and complement more common methods such as differential gene expression, thus providing a more complete picture of the biological differences across samples or groups.

References

Adil, Asif, Vijay Kumar, Arif Tasleem Jan, and Mohammed Asger. 2021. “Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis.” *Frontiers in Neuroscience* 15 (April): 591122.

Aebi, Markus. 2013. “N-Linked Protein Glycosylation in the ER.” *Biochimica et Biophysica Acta* 1833 (11): 2430–37.

Agren, Rasmus, Sergio Bordel, Adil Mardinoglu, Natapol Pornputtapong, Intawat Nookaew, and Jens Nielsen. 2012. “Reconstruction of Genome-Scale Active Metabolic Networks for 69 Human Cell Types and 16 Cancer Types Using INIT.” *PLoS Computational Biology* 8 (5): e1002518.

Agren, Rasmus, Adil Mardinoglu, Anna Asplund, Caroline Kampf, Mathias Uhlen, and Jens Nielsen. 2014. “Identification of Anticancer Drugs for Hepatocellular Carcinoma through Personalized Genome-Scale Metabolic Modeling.” *Molecular Systems Biology* 10 (3): 721.

Aiba, Yoshihiro, Kenichi Harada, Masahiro Ito, Takashi Suematsu, Shinichi Aishima, Yuki Hitomi, Nao Nishida, et al. 2018. “Increased Expression and Altered Localization of Cathepsin Z Are Associated with Progression to Jaundice Stage in Primary Biliary Cholangitis.” *Scientific Reports* 8 (1): 11808.

Akiyama, Masato. 2021. “Multi-Omics Study for Interpretation of Genome-Wide Association Study.” *Journal of Human Genetics* 66 (1): 3–10.

Alberts, Bruce. 2017. *Molecular Biology of the Cell*. Garland Science.

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002a. *Transport from the ER through the Golgi Apparatus*. Garland Science.

———. 2002b. *Transport from the Trans Golgi Network to the Cell Exterior: Exocytosis*. Garland Science.

Aldridge, Sarah, and Sarah A. Teichmann. 2020. “Single Cell Transcriptomics Comes of Age.” *Nature Communications* 11 (1): 4307.

Alexander, Lisa M., Daniel H. Goldman, Liang M. Wee, and Carlos Bustamante. 2019. “Non-Equilibrium Dynamics of a Nascent Polypeptide during Translation Suppress Its Misfolding.” *Nature Communications* 10 (1): 2709.

Andrews, Tallulah S., Vladimir Yu Kiselev, Davis McCarthy, and Martin Hemberg. 2021. “Tutorial: Guidelines for the Computational Analysis of Single-Cell RNA Sequencing Data.” *Nature Protocols* 16 (1): 1–9.

An, Hyun Joo, John W. Froehlich, and Carlito B. Lebrilla. 2009. “Determination of Glycosylation Sites and Site-Specific Heterogeneity in Glycoproteins.” *Current Opinion in Chemical Biology* 13 (4): 421–26.

Ankeny, Rachel A. 2003. “Sequencing the Genome from Nematode to Human: Changing Methods, Changing Science.” *Endeavour* 27 (2): 87–92.

Araki, Kazutaka, and Kazuhiro Nagata. 2011. “Protein Folding and Quality Control in the ER.” *Cold Spring Harbor Perspectives in Biology* 3 (11): a007526.

Argelaguet, Ricard, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. 2018. “Multi-Omics Factor Analysis-a Framework for Unsupervised Integration of Multi-Omics Data Sets.” *Molecular Systems Biology* 14 (6): e8124.

Aviram, Naama, and Maya Schuldiner. 2017. “Targeting and Translocation of Proteins to the Endoplasmic Reticulum at a Glance.” *Journal of Cell Science* 130 (24): 4079–85.

Baart, Gino J. E., and Dirk E. Martens. 2012. “Genome-Scale Metabolic Models: Reconstruction and Analysis.” *Methods in Molecular Biology* 799: 107–26.

Baeshen, Nabih A., Mohammed N. Baeshen, Abdullah Sheikh, Roop S. Bora, Mohamed Morsi M. Ahmed, Hassan A. I. Ramadan, Kulvinder Singh Saini, and Elrashdy M. Redwan. 2014. “Cell Factories for Insulin Production.” *Microbial Cell Factories* 13 (October): 141.

Baron, Ryan D., and Francis A. Barr. 2015. “The Kinesin-6 Members MKLP1, MKLP2 and MPP1.” In *Kinesins and Cancer*, edited by Kozielski, FSB, and Frank, 193–222. Dordrecht: Springer Netherlands.

Behrouz, Hossein, Behnaz Molavi, Ata Tavakoli, Mansoureh Askari, Shayan Maleknia, Fereidoun Mahboudi, and Mehdi Khodadadian. 2020. “Multivariate Optimization of the Refolding Process of an Incorrectly Folded Fc-Fusion Protein in a Cell Culture Broth.” *Current Pharmaceutical Biotechnology* 21 (3): 226–35.

Beissinger, M., and J. Buchner. 1998. “How Chaperones Fold Proteins.” *Biological Chemistry* 379 (3): 245–59.

Berg, Jeremy M., John L. Tymoczko, and Lubert Stryer. 2002. *Cells Can Respond to Changes in Their Environments*. W H Freeman.

Berlec, Aleš, and Borut Strukelj. 2013. “Current State and Recent Advances in Biopharmaceutical Production in Escherichia Coli, Yeasts and Mammalian Cells.” *Journal of Industrial Microbiology & Biotechnology* 40 (3-4): 257–74.

Bersanelli, Matteo, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. 2016. “Methods for the Integration of Multi-Omics Data: Mathematical Aspects.” *BMC Bioinformatics* 17 Suppl 2 (January): 15.

Bhalla, U. S., and R. Iyengar. 1999. “Emergent Properties of Networks of Biological Signaling Pathways.” *Science* 283 (5400): 381–87.

Blum, Benjamin C., Fatemeh Mousavi, and Andrew Emili. 2018. “Single-Platform ‘Multi-Omic’ Profiling: Unified Mass Spectrometry and Computational Workflows for Integrative Proteomics-Metabolomics Analysis.” *Molecular Omics* 14 (5): 307–19.

Boccard, Julien, and Serge Rudaz. 2016. “Exploring Omics Data from Designed Experiments Using Analysis of Variance Multiblock Orthogonal Partial Least Squares.” *Analytica Chimica Acta* 920 (May): 18–28.

Bogdanovic, Nenad, Oskar Hansson, Henrik Zetterberg, Hans Basun, Martin Ingelsson, Lars Lannfelt, and Kaj Blennow. 2020. “[Alzheimer’s disease - the most common cause of dementia].” *Lakartidningen* 117 (March). https://www.ncbi.nlm.nih.gov/pubmed/32154904.

Böhm, Ernst, Birgit K. Seyfried, Michael Dockal, Michael Graninger, Meinhard Hasslacher, Marianne Neurath, Christian Konetschny, Peter Matthiessen, Artur Mitterer, and Friedrich Scheiflinger. 2015. “Differences in N-Glycosylation of Recombinant Human Coagulation Factor VII Derived from BHK, CHO, and HEK293 Cells.” *BMC Biotechnology* 15 (September): 87.

Bordbar, Aarash, Jonathan M. Monk, Zachary A. King, and Bernhard O. Palsson. 2014. “Constraint-Based Models Predict Metabolic and Associated Cellular Functions.” *Nature Reviews. Genetics* 15 (2): 107–20.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. “A Training Algorithm for Optimal Margin Classifiers.” In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–52. COLT ’92. New York, NY, USA: Association for Computing Machinery.

Braakman, Ineke, and Daniel N. Hebert. 2013. “Protein Folding in the Endoplasmic Reticulum.” *Cold Spring Harbor Perspectives in Biology* 5 (5): a013201.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Brosnan, John T., and Margaret E. Brosnan. 2010. “Creatine Metabolism and the Urea Cycle.” *Molecular Genetics and Metabolism* 100 Suppl 1 (March): S49–52.

Brown, D., and S. Breton. 2000. “Sorting Proteins to Their Target Membranes.” *Kidney International* 57 (3): 816–24.

Brunk, Elizabeth, Swagatika Sahoo, Daniel C. Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, et al. 2018. “Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism.” *Nature Biotechnology* 36 (3): 272–81.

Bruno, Benjamin J., Geoffrey D. Miller, and Carol S. Lim. 2013. “Basics and Recent Advances in Peptide and Protein Drug Delivery.” *Therapeutic Delivery* 4 (11): 1443–67.

Burda, Patricie, and Markus Aebi. 1999. “The Dolichol Pathway of N-Linked Glycosylation.” *Biochimica et Biophysica Acta - General Subjects*. https://doi.org/10.1016/S0304-4165(98)00127-5.

Butler, Michael, and Maureen Spearman. 2014. “The Choice of Mammalian Cell Host and Possibilities for Glycosylation Engineering.” *Current Opinion in Biotechnology* 30 (December): 107–12.

Caspi, Ron, Hartmut Foerster, Carol A. Fulcher, Pallavi Kaipa, Markus Krummenacker, Mario Latendresse, Suzanne Paley, et al. 2008. “The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases.” *Nucleic Acids Research* 36 (Database issue): D623–31.

Chang, T. W. 1983. “Binding of Cells to Matrixes of Distinct Antibodies Coated on Solid Surface.” *Journal of Immunological Methods* 65 (1-2): 217–23.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD ’16. New York, NY, USA: Association for Computing Machinery.

Chen, Yiqun, Brian O. McConnell, Venkata Gayatri Dhara, Harnish Mukesh Naik, Chien-Ting Li, Maciek R. Antoniewicz, and Michael J. Betenbaugh. 2019. “An Unconventional Uptake Rate Objective Function Approach Enhances Applicability of Genome-Scale Models for Mammalian Cells.” *NPJ Systems Biology and Applications* 5 (July): 25.

Chen, Yu, and Jens Nielsen. 2019. “Energy Metabolism Controls Phenotypes by Protein Efficiency and Allocation.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (35): 17592–97.

———. 2021. “Mathematical Modeling of Proteome Constraints within Metabolism.” *Current Opinion in Systems Biology*. https://doi.org/10.1016/j.coisb.2021.03.003.

Chen, Yu, Jens Nielsen, and Eduard J. Kerkhoven. 2021. “Proteome Constraints in Genome-scale Models.” *Metabolic Engineering*. Wiley. https://doi.org/10.1002/9783527823468.ch4.

Church, George M. 2006. “Genomes for All.” *Scientific American* 294 (1): 46–54.

Chu, Yongjun, and David R. Corey. 2012. “RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation.” *Nucleic Acid Therapeutics* 22 (4): 271–74.
Cohen, S. N., A. C. Chang, H. W. Boyer, and R. B. Helling. 1973. “Construction of Biologically Functional Bacterial Plasmids in Vitro.” *Proceedings of the National Academy of Sciences of the United States of America* 70 (11): 3240–44.
Cong, Le, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, et al. 2013. “Multiplex Genome Engineering Using CRISPR/Cas Systems.” *Science* 339 (6121): 819–23.
Costa, Cristine Alves da, Wejdane El Manaa, Eric Duplan, and Frédéric Checler. 2020. “The Endoplasmic Reticulum Stress/Unfolded Protein Response and Their Contributions to Parkinson’s Disease Physiopathology.” *Cells* 9 (11). https://doi.org/10.3390/cells9112495.
Côté, Johanne, Alain Garnier, Bernard Massie, and Amine Kamen. 1998. “Serum-Free Production of Recombinant Proteins and Adenoviral Vectors by 293SF-3F6 Cells.” *Biotechnology and Bioengineering* 59 (5): 567–75.
Craig-Mueller, Nils, Ruba Hammad, Roland Elling, Jamal Alzubi, Barbara Timm, Julia Kolter, Nele Knelangen, et al. 2020. “Modeling MyD88 Deficiency In Vitro Provides New Insights in Its Function.” *Frontiers in Immunology* 11 (December): 608802.
Cummings, Richard D., and J. Michael Pierce. 2014. “The Challenge and Promise of Glycomics.” *Chemistry & Biology* 21 (1): 1–15.
Cupp-Sutton, Kellye A., and Si Wu. 2020. “High-Throughput Quantitative Top-down Proteomics.” *Molecular Omics* 16 (2): 91–99.
Dahal, Sanjeev, James T. Yurkovich, Hao Xu, Bernhard O. Palsson, and Laurence Yang. 2020. “Synthesizing Systems Biology Knowledge from Omics Using Genome-Scale Models.” *Proteomics* 20 (17-18): e1900282.
Das, Tonmoy, Geoffroy Andrieux, Musaddeque Ahmed, and Sajib Chakraborty. 2020. “Integration of Online Omics-Data Resources for Cancer Research.” *Frontiers in Genetics* 11 (October): 578345.
Del Val, Ioscani Jimenez, Karen M. Polizzi, and Cleo Kontoravdi. 2016. “A Theoretical Estimate for Nucleotide Sugar Demand towards Chinese Hamster Ovary Cellular Glycosylation.” *Scientific Reports* 6 (June): 28547.
Diez-Fernandez, Carmen, and Johannes Häberle. 2017. “Targeting CPS1 in the Treatment of Carbamoyl Phosphate Synthetase 1 (CPS1) Deficiency, a Urea Cycle Disorder.” *Expert Opinion on Therapeutic Targets* 21 (4): 391–99.
Di Girolamo, Francesco, Isabella Lante, Maurizio Muraca, and Lorenza Putignani. 2013. “The Role of Mass Spectrometry in the ‘Omics’ Era.” *Current Organic Chemistry* 17 (23): 2891–2905.
Ding, Mei, Hanna Tegel, Åsa Sivertsson, Sophia Hober, Arjan Snijder, Mats Ormö, Per-Erik Strömstedt, Rick Davies, and Lovisa Holmberg Schiavone. 2020. “Secretome-Based Screening in Target Discovery.” *SLAS Discovery : Advancing Life Sciences R & D* 25 (6): 535–51.
Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.
Domenzain, I., B. Sánchez, M. Anton, and E. J. Kerkhoven. 2021. “Reconstruction of a Catalogue of Genome-Scale Metabolic Models with Enzymatic Constraints Using GECKO 2.0.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.03.05.433259v1.abstract.
Domingos, Pedro, and Michael Pazzani. 1997. “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss.” *Machine Learning* 29 (2): 103–30.
Drăghici, Sorin, and R. Brian Potter. 2003. “Predicting HIV Drug Resistance with Neural Networks.” *Bioinformatics* 19 (1): 98–107.
Duan, Guangyou, and Dirk Walther. 2015. “The Roles of Post-Translational Modifications in the Context of Protein Interaction Networks.” *PLoS Computational Biology* 11 (2): e1004049.
Duarte, Natalie C., Scott A. Becker, Neema Jamshidi, Ines Thiele, Monica L. Mo, Thuy D. Vo, Rohith Srivas, and Bernhard Ø. Palsson. 2007. “Global Reconstruction of the Human Metabolic Network Based on Genomic and Bibliomic Data.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (6): 1777–82.
Du, Bin, Laurence Yang, Colton J. Lloyd, Xin Fang, and Bernhard O. Palsson. 2019. “Genome-Scale Model of Metabolism and Gene Expression Provides a Multi-Scale Description of Acid Stress Responses in Escherichia Coli.” *PLoS Computational Biology* 15 (12): e1007525.
Dugger, Brittany N., and Dennis W. Dickson. 2017. “Pathology of Neurodegenerative Diseases.” *Cold Spring Harbor Perspectives in Biology* 9 (7). https://doi.org/10.1101/cshperspect.a028035.
Dumont, Jennifer, Don Euwart, Baisong Mei, Scott Estes, and Rashmi Kshirsagar. 2016. “Human Cell Lines for Biopharmaceutical Manufacturing: History, Status, and Future Perspectives.” *Critical Reviews in Biotechnology* 36 (6): 1110–22.
Dunn, William A. 2003. “Pathways of Mammalian Protein Degradation.” In *New Comprehensive*

Biochemistry, 38:513–33. Elsevier.
Edwards, J. S., and B. O. Palsson. 1999. “Systems Properties of the Haemophilus Influenzae Rd Metabolic Genotype.” *The Journal of Biological Chemistry* 274 (25): 17410–16.
Eldeeb, Mohamed A., Ramanaguru Siva-Piragasam, Mohamed A. Ragheb, Mansoore Esmaili, Mohamed Salla, and Richard P. Fahlman. 2019. “A Molecular Toolbox for Studying Protein Degradation in Mammalian Cells.” *Journal of Neurochemistry* 151 (4): 520–33.
Ellegren, Hans. 2008. “Comparative Genomics and the Study of Evolution by Natural Selection.” *Molecular Ecology* 17 (21): 4586–96.
Ellgaard, Lars, Nicholas McCaul, Anna Chatsisvili, and Ineke Braakman. 2016. “Co- and Post-Translational Protein Folding in the ER.” *Traffic*. https://doi.org/10.1111/tra.12392.
Ellgaard, L., and A. Helenius. 2001. “ER Quality Control: Towards an Understanding at the Molecular Level.” *Current Opinion in Cell Biology* 13 (4): 431–37.
Fang, Xin, Colton J. Lloyd, and Bernhard O. Palsson. 2020. “Reconstructing Organisms in Silico: Genome-Scale Models and Their Emerging Applications.” *Nature Reviews. Microbiology*, September. https://doi.org/10.1038/s41579-020-00440-4.
Faye, L., A. Sturm, R. Bollini, A. Vitale, and M. J. Chrispeels. 1986. “The Position of the Oligosaccharide Side-Chains of Phytohemagglutinin and Their Accessibility to Glycosidases Determines Their Subsequent Processing in the Golgi.” *European Journal of Biochemistry / FEBS* 158 (3): 655–61.
Feizi, Amir, Amir Banaei-Esfahani, and Jens Nielsen. 2015. “HCSD: The Human Cancer Secretome Database.” *Database: The Journal of Biological Databases and Curation* 2015 (June): bav051.
Feizi, Amir, Francesco Gatto, Mathias Uhlen, and Jens Nielsen. 2017. “Human Protein Secretory Pathway Genes Are Expressed in a Tissue-Specific Pattern to Match Processing Demands of the Secretome.” *NPJ Systems Biology and Applications* 3 (August): 22.
Feizi, Amir, Tobias Österlund, Dina Petranovic, Sergio Bordel, and Jens Nielsen. 2013. “Genome-Scale Modeling of the Protein Secretory Machinery in Yeast.” *PloS One* 8 (5): e63284.
Ferreira, I. L., R. Resende, E. Ferreiro, A. C. Rego, and C. F. Pereira. 2010. “Multiple Defects in Energy Metabolism in Alzheimer’s Disease.” *Current Drug Targets* 11 (10): 1193–1206.
Fewell, Sheara W., and Jeffrey L. Brodsky. 2013. *Entry into the Endoplasmic Reticulum: Protein Translocation, Folding and Quality Control*. Landes Bioscience.
Fischer, M. 2017. “Census and Evaluation of p53 Target Genes.” *Oncogene* 36 (28): 3943–56.
Fischer, Martin, Inga Grundke, Sindy Sohr, Marianne Quaas, Saskia Hoffmann, Arne Knörck, Catalina Gumhold, and Karen Rother. 2013. “p53 and Cell Cycle Dependent Transcription of Kinesin Family Member 23 (KIF23) Is Controlled via a CHR Promoter Element Bound by DREAM and MMB Complexes.” *PloS One* 8 (5): e63187.
Fischer, Simon, René Handrick, and Kerstin Otte. 2015. “The Art of CHO Cell Engineering: A Comprehensive Retrospect and Future Perspectives.” *Biotechnology Advances* 33 (8): 1878–96.
Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. 1995. “Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd.” *Science* 269 (5223): 496–512.
Flicek, Paul, Ikhlak Ahmed, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, et al. 2013. “Ensembl 2013.” *Nucleic Acids Research* 41 (Database issue): D48–55.
Fouladiha, Hamideh, Sayed-Amir Marashi, Shangzhong Li, Zerong Li, Helen O. Masson, Behrouz Vaziri, and Nathan E. Lewis. 2021. “Systematically Gap-Filling the Genome-Scale Metabolic Model of CHO Cells.” *Biotechnology Letters* 43 (1): 73–87.
Freund, Yoav, and Robert E. Schapire. 1997. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” *Journal of Computer and System Sciences* 55 (1): 119–39.
Garnier, A., J. Côté, I. Nadeau, A. Kamen, and B. Massie. 1994. “Scale-up of the Adenovirus Expression System for the Production of Recombinant Protein in Human 293S Cells.” *Cytotechnology* 15 (1-3): 145–55.
Gasnereau, Isabelle, Mathieu Boissan, Germain Margall-Ducos, Gabrielle Couchy, Dominique Wendum, Florence Bourgain-Guglielmetti, Chantal Desdouets, Marie-Lise Lacombe, Jessica Zucman-Rossi, and Joëlle Sobczak-Thépot. 2012. “KIF20A mRNA and Its Product MKlp2 Are Increased during Hepatocyte Proliferation and Hepatocarcinogenesis.” *The American Journal of Pathology* 180 (1): 131–40.
Genuth, Naomi R., and Maria Barna. 2018. “The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life.” *Molecular Cell* 71 (3): 364–74.
Geurts, Pierre, Damien Ernst, and Louis Wehenkel. 2006. “Extremely Randomized Trees.” *Machine Learning* 63 (1): 3–42.
Giuliani, Maria, Ermenegilda Parrilli, Pau Ferrer, Kristin Baumann, Gennaro Marino, and Maria Luisa Tutino. 2011. “Process Optimization for Recombinant Protein Production in the Psychrophilic Bacterium Pseudoalteromonas Haloplanktis.” *Process Biochemistry* 46 (4): 953–59.

Glick, B. S. 2000. “Organization of the Golgi Apparatus.” *Current Opinion in Cell Biology* 12 (4): 450–56.
Gogala, Marko, Thomas Becker, Birgitta Beatrix, Jean-Paul Armache, Clara Barrio-Garcia, Otto Berninghausen, and Roland Beckmann. 2014. “Structures of the Sec61 Complex Engaged in Nascent Peptide Translocation or Membrane Insertion.” *Nature* 506 (7486): 107–10.
Goh, Justin Bryan, and Say Kong Ng. 2018. “Impact of Host Cell Line Choice on Glycan Profile.” *Critical Reviews in Biotechnology* 38 (6): 851–67.
Gomez-Navarro, Natalia, and Elizabeth Miller. 2016. “Protein Sorting at the ER–Golgi Interface.” *The Journal of Cell Biology* 215 (6): 769–78.
Gopal Krishnan, Priya D., Emily Golden, Eleanor A. Woodward, Nathan J. Pavlos, and Pilar Blancafort. 2020. “Rab GTPases: Emerging Oncogenes and Tumor Suppressive Regulators for the Editing of Survival Pathways in Cancer.” *Cancers* 12 (2). https://doi.org/10.3390/cancers12020259.
Goto, Masatoshi. 2007. “Protein O-Glycosylation in Fungi: Diverse Structures and Multiple Functions.” *Bioscience, Biotechnology, and Biochemistry* 71 (6): 1415–27.
Graham, F. L. 1987. “Growth of 293 Cells in Suspension Culture.” *The Journal of General Virology* 68 (3): 937–40.
Graham, F. L., J. Smiley, W. C. Russell, and R. Nairn. 1977. “Characteristics of a Human Cell Line Transformed by DNA from Human Adenovirus Type 5.” *The Journal of General Virology* 36 (1): 59–74.
Graupner, V., E. Alexander, T. Overkamp, O. Rothfuss, V. De Laurenzi, B. F. Gillissen, P. T. Daniel, K. Schulze-Osthoff, and F. Essmann. 2011. “Differential Regulation of the Proapoptotic Multidomain Protein Bak by p53 and p73 at the Promoter Level.” *Cell Death and Differentiation* 18 (7): 1130–39.
Gu, Changdai, Gi Bae Kim, Won Jun Kim, Hyun Uk Kim, and Sang Yup Lee. 2019. “Current Status and Applications of Genome-Scale Metabolic Models.” *Genome Biology* 20 (1): 121.
Guo, Yusong, Daniel W. Sirkis, and Randy Schekman. 2014. “Protein Sorting at the Trans-Golgi Network.” *Annual Review of Cell and Developmental Biology* 30 (August): 169–206.
Gupta, Yashdeep, Gaurav Singla, and Rajiv Singla. 2015. “Insulin-Derived Amyloidosis.” *Indian Journal of Endocrinology and Metabolism* 19 (1): 174–77.
Gutierrez, Jahir M., Amir Feizi, Shangzhong Li, Thomas B. Kallehauge, Hooman Hefzi, Lise M. Grav, Daniel Ley, et al. 2020. “Genome-Scale Reconstructions of the Mammalian Secretory Pathway Predict Metabolic Costs and Limitations of Protein Secretion.” *Nature Communications* 11 (1): 68.
Gygi, S. P., Y. Rochon, B. R. Franza, and R. Aebersold. 1999. “Correlation between Protein and mRNA Abundance in Yeast.” *Molecular and Cellular Biology* 19 (3): 1720–30.
Hadadi, Noushin, Vikash Pandey, Anush Chiappino-Pepe, Marian Morales, Hector Gallart-Ayala, Florence Mehl, Julijana Ivanisevic, Vladimir Sentchilo, and Jan R. van der Meer. 2020. “Mechanistic Insights into Bacterial Metabolic Reprogramming from Omics-Integrated Genome-Scale Models.” *NPJ Systems Biology and Applications* 6 (1): 1.
Hall, Neil. 2007. “Advanced Sequencing Technologies and Their Wider Impact in Microbiology.” *The Journal of Experimental Biology* 210 (Pt 9): 1518–25.
“Handbook of Statistics.” n.d. Accessed June 9, 2021. https://www.sciencedirect.com/handbook/handbook-of-statistics.
Hanson, Sarah R., Elizabeth K. Culyba, T-L Hsu, Chi-Huey Wong, Jeffery W. Kelly, and Evan T. Powers. 2009. “The Core Trisaccharide of an N-Linked Glycoprotein Intrinsically Accelerates Folding and Enhances Stability.” *Proceedings of the National Academy of Sciences* 106 (9): 3131–36.
Hardison, Ross C. 2003. “Comparative Genomics.” *PLoS Biology* 1 (2): E58.
Hastings, Janna, Abraham Mains, Bhupinder Virk, Nicolas Rodriguez, Sharlene Murdoch, Juliette Pearce, Sven Bergmann, Nicolas Le Novère, and Olivia Casanueva. 2019. “Multi-Omics and Genome-Scale Modeling Reveal a Metabolic Shift During C. Elegans Aging.” *Frontiers in Molecular Biosciences* 6 (February): 2.
Hegde, Ramanujan S., and Robert J. Keenan. 2011. “Tail-Anchored Membrane Protein Insertion into the Endoplasmic Reticulum.” *Nature Reviews. Molecular Cell Biology* 12 (12): 787–98.
Helenius, A., and M. Aebi. 2001. “Intracellular Functions of N-Linked Glycans.” *Science* 291 (5512): 2364–69.
“Hematology.” n.d. Accessed May 3, 2021. https://www.sciencedirect.com/book/9780323357623/hematology.
Hochstrasser, M. 1996. “Ubiquitin-Dependent Protein Degradation.” *Annual Review of Genetics* 30: 405–39.
Ho, Tin Kam. 1995. “Random Decision Forests.” In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–82 vol.1.
Hsu, Victor W., Stella Y. Lee, and Jia-Shu Yang. 2009. “The Evolving Understanding of COPI Vesicle Formation.” *Nature Reviews. Molecular Cell Biology* 10 (5): 360–64.
Huang, Mingtao, Jichen Bao, Björn M. Hallström, Dina Petranovic, and Jens Nielsen. 2017. “Efficient

Protein Production by Yeast Requires Global Tuning of Metabolism." *Nature Communications* 8 (1): 1131.
Huang, Sijia, Kumardeep Chaudhary, and Lana X. Garmire. 2017. "More Is Better: Recent Progress in Multi-Omics Data Integration Methods." *Frontiers in Genetics* 8 (June): 84.
Huertas, Maria José, and Carmen Michán. 2019. "Paving the Way for the Production of Secretory Proteins by Yeast Cell Factories." *Microbial Biotechnology* 12 (6): 1095–96.
Hung, Mien-Chie, and Wolfgang Link. 2011. "Protein Localization in Disease and Therapy." *Journal of Cell Science* 124 (Pt 20): 3381–92.
Huntley, Stuart, Daniel M. Baggott, Aaron T. Hamilton, Mary Tran-Gyamfi, Shan Yang, Joomyeong Kim, Laurie Gordon, Elbert Branscomb, and Lisa Stubbs. 2006. "A Comprehensive Catalog of Human KRAB-Associated Zinc Finger Genes: Insights into the Evolutionary History of a Large Family of Transcriptional Repressors." *Genome Research* 16 (5): 669–77.
Hunt, Sylvie D., and David J. Stephens. 2011. "The Role of Motor Proteins in Endosomal Sorting." *Biochemical Society Transactions* 39 (5): 1179–84.
Hyduke, Daniel R., Nathan E. Lewis, and Bernhard Ø. Palsson. 2013. "Analysis of Omics Data with Genome-Scale Models of Metabolism." *Molecular bioSystems* 9 (2): 167–74.
Isermann, Anna, Carl Mann, and Claudia E. Rübe. 2020. "Histone Variant H2A.J Marks Persistent DNA Damage and Triggers the Secretory Phenotype in Radiation-Induced Senescence." *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms21239130.
Jahn, Reinhard, and Richard H. Scheller. 2006. "SNAREs--Engines for Membrane Fusion." *Nature Reviews. Molecular Cell Biology* 7 (9): 631–43.
Jenkins, Nigel. 2007. "Modifications of Therapeutic Proteins: Challenges and Prospects." *Cytotechnology* 53 (1-3): 121–25.
Jenkins, Nigel, Lisa Murphy, and Ray Tyther. 2008. "Post-Translational Modifications of Recombinant Proteins: Significance for Biopharmaceuticals." *Molecular Biotechnology* 39 (2): 113–18.
Jensen, Devon, and Randy Schekman. 2011. "COPII-Mediated Vesicle Formation at a Glance." *Journal of Cell Science* 124 (Pt 1): 1–4.
Jensen-Jarolim, Erika, ed. 2014. *Comparative Medicine: Anatomy and Physiology*. Springer, Vienna.
Johnson, Nicholas, Katie Powis, and Stephen High. 2013. "Post-Translational Translocation into the Endoplasmic Reticulum." *Biochimica et Biophysica Acta* 1833 (11): 2403–9.
Jolliffe, Ian T., and Jorge Cadima. 2016. "Principal Component Analysis: A Review and Recent Developments." *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374 (2065): 20150202.
Jucker, Mathias, and Lary C. Walker. 2013. "Self-Propagation of Pathogenic Protein Aggregates in Neurodegenerative Diseases." *Nature* 501 (7465): 45–51.
Kamisoglu, Kubra, Alison Acevedo, Richard R. Almon, Susette Coyle, Siobhan Corbett, Debra C. Dubois, Tung T. Nguyen, William J. Jusko, and Ioannis P. Androulakis. 2017. "Understanding Physiology in the Continuum: Integration of Information from Multiple -Omics Levels." *Frontiers in Pharmacology* 8 (February): 91.
Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. 2021. "KEGG: Integrating Viruses and Cellular Organisms." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa970.
Kawaguchi, Shinichi, and Davis T. W. Ng. 2011. "Sensing ER Stress." *Science* 333 (6051): 1830–31.
Khanabdali, Ramin, Ayeshah A. Rosdah, Gregory J. Dusting, and Shiang Y. Lim. 2016. "Harnessing the Secretome of Cardiac Stem Cells as Therapy for Ischemic Heart Disease." *Biochemical Pharmacology* 113 (August): 1–11.
Kim, Jee Yon, Yeon-Gu Kim, and Gyun Min Lee. 2012. "CHO Cells in Biotechnology for Production of Recombinant Proteins: Current State and Further Potential." *Applied Microbiology and Biotechnology* 93 (3): 917–30.
Kintzing, James R., Maria V. Filsinger Interrante, and Jennifer R. Cochran. 2016. "Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment." *Trends in Pharmacological Sciences* 37 (12): 993–1008.
Kol, Stefan, Daniel Ley, Tune Wulff, Marianne Decker, Johnny Arnsdorf, Sanne Schoffelen, Anders Holmgaard Hansen, et al. 2020. "Multiplex Secretome Engineering Enhances Recombinant Protein Production and Purity." *Nature Communications* 11 (1): 1908.
Kweon, Jiyeon, and Yongsub Kim. 2018. "High-Throughput Genetic Screens Using CRISPR-Cas9 System." *Archives of Pharmacal Research* 41 (9): 875–84.
Lai, F., A. A. Fernald, N. Zhao, and M. M. Le Beau. 2000. "cDNA Cloning, Expression Pattern, Genomic Structure and Chromosomal Location of RAB6KIFL, a Human Kinesin-like Gene." *Gene* 248 (1-2): 117–25.
Lancaster, Samuel M., Akshay Sanghi, Si Wu, and Michael P. Snyder. 2020. "A Customizable Analysis

Flow in Integrative Multi-Omics." *Biomolecules* 10 (12). https://doi.org/10.3390/biom10121606.
Lau, Jolene L., and Michael K. Dunn. 2018. "Therapeutic Peptides: Historical Perspectives, Current Development Trends, and Future Directions." *Bioorganic & Medicinal Chemistry* 26 (10): 2700–2707.
Lebeaupin, Cynthia, Jing Yong, and Randal J. Kaufman. 2020. "The Impact of the ER Unfolded Protein Response on Cancer Initiation and Progression: Therapeutic Implications." *Advances in Experimental Medicine and Biology* 1243: 113–31.
Le Fourn, Valérie, Pierre-Alain Girod, Montse Buceta, Alexandre Regamey, and Nicolas Mermod. 2014. "CHO Cell Engineering to Prevent Polypeptide Aggregation and Improve Therapeutic Protein Secretion." *Metabolic Engineering* 21 (January): 91–102.
Lenk, Uwe, Helen Yu, Jan Walter, Marina S. Gelman, Enno Hartmann, Ron R. Kopito, and Thomas Sommer. 2002. "A Role for Mammalian Ubc6 Homologues in ER-Associated Protein Degradation." *Journal of Cell Science* 115 (Pt 14): 3007–14.
Lerman, Joshua A., Daniel R. Hyduke, Haythem Latif, Vasiliy A. Portnoy, Nathan E. Lewis, Jeffrey D. Orth, Alexandra C. Schrimpe-Rutledge, et al. 2012. "In Silico Method for Modelling Metabolism and Gene Product Expression at Genome Scale." *Nature Communications* 3 (July): 929.
Levenson, Robert W., Virginia E. Sturm, and Claudia M. Haase. 2014. "Emotional and Behavioral Symptoms in Neurodegenerative Disease: A Model for Studying the Neural Bases of Psychopathology." *Annual Review of Clinical Psychology* 10 (January): 581–606.
Lewis, Nathan E., Harish Nagarajan, and Bernhard O. Palsson. 2012. "Constraining the Metabolic Genotype–phenotype Relationship Using a Phylogeny of in Silico Methods." *Nature Reviews. Microbiology* 10 (4): 291–305.
Liang, Chenguang, Austin W. T. Chiang, Anders H. Hansen, Johnny Arnsdorf, Sanne Schoffelen, James T. Sorrentino, Benjamin P. Kellman, Bokan Bao, Bjørn G. Voldborg, and Nathan E. Lewis. 2020. "A Markov Model of Glycosylation Elucidates Isozyme Specificity and Glycosyltransferase Interactions for Glycoengineering." *Current Research in Biotechnology* 2 (November): 22–36.
Li, Dongmei. 2019. "Statistical Methods for RNA Sequencing Data Analysis." In *Computational Biology*, edited by Holger Husi. Brisbane (AU): Codon Publications.
Lin, Dongdong, Hima B. Yalamanchili, Xinmin Zhang, Nathan E. Lewis, Christina S. Alves, Joost Groot, Johnny Arnsdorf, et al. 2020. "CHOmics: A Web-Based Tool for Multi-Omics Data Analysis and Interactive Visualization in CHO Cell Lines." *PLoS Computational Biology* 16 (12): e1008498.
Lin, Yao-Cheng, Morgane Boone, Leander Meuris, Irma Lemmens, Nadine Van Roy, Arne Soete, Joke Reumers, et al. 2014. "Genome Dynamics of the Human Embryonic Kidney 293 Lineage in Response to Cell Biology Manipulations." *Nature Communications* 5 (September): 4767.
Liste-Calleja, Leticia, Martí Lecina, and Jordi Joan Cairó. 2013. "HEK293 Cell Culture Media Study: Increasing Cell Density for Different Bioprocess Applications." *BMC Proceedings* 7 (6): P51.
Liu, Ming, Michael A. Weiss, Anoop Arunagiri, Jing Yong, Nischay Rege, Jinhong Sun, Leena Haataja, Randal J. Kaufman, and Peter Arvan. 2018. "Biosynthesis, Structure, and Folding of the Insulin Precursor Protein." *Diabetes, Obesity & Metabolism* 20 Suppl 2 (September): 28–50.
Lloyd, Colton J., Ali Ebrahim, Laurence Yang, Zachary A. King, Edward Catoiu, Edward J. O'Brien, Joanne K. Liu, and Bernhard O. Palsson. 2018. "COBRAme: A Computational Framework for Genome-Scale Models of Metabolism and Gene Expression." *PLoS Computational Biology* 14 (7): e1006302.
Lodish, Harvey, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. 2000. *Overview of the Secretory Pathway*. W. H. Freeman.
Loira, Nicolas, Thierry Dulermo, Jean-Marc Nicaud, and David James Sherman. 2012. "A Genome-Scale Metabolic Model of the Lipid-Accumulating Yeast Yarrowia Lipolytica." *BMC Systems Biology* 6 (May): 35.
Long, Justin M., and David M. Holtzman. 2019. "Alzheimer Disease: An Update on Pathobiology and Treatment Strategies." *Cell* 179 (2): 312–39.
López de Maturana, Evangelina, Lola Alonso, Pablo Alarcón, Isabel Adoración Martín-Antoniano, Silvia Pineda, Lucas Piorno, M. Luz Calle, and Núria Malats. 2019. "Challenges in the Integration of Omics and Non-Omics Data." *Genes* 10 (3). https://doi.org/10.3390/genes10030238.
Losev, Eugene, Catherine A. Reinke, Jennifer Jellen, Daniel E. Strongin, Brooke J. Bevis, and Benjamin S. Glick. 2006. "Golgi Maturation Visualized in Living Yeast." *Nature* 441 (7096): 1002–6.
Louis, Nathalie, Carole Evelegh, and Frank L. Graham. 1997. "Cloning and Sequencing of the Cellular–Viral Junctions from the Human Adenovirus Type 5 Transformed 293 Cell Line." *Virology* 233 (2): 423–29.
Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.
Lowenthal, Mark S., Kiersta S. Davis, Trina Formolo, Lisa E. Kilpatrick, and Karen W. Phinney. 2016. "Identification of Novel N-Glycosylation Sites at Noncanonical Protein Consensus Motifs." *Journal of*

Proteome Research 15 (7): 2087–2101.
Luheshi, Leila M., Damian C. Crowther, and Christopher M. Dobson. 2008. “Protein Misfolding and Disease: From the Test Tube to the Organism.” *Current Opinion in Chemical Biology* 12 (1): 25–31.
Mahlab, Shelly, and Michal Linial. 2014. “Speed Controls in Translating Secretory Proteins in Eukaryotes--an Evolutionary Perspective.” *PLoS Computational Biology* 10 (1): e1003294.
Ma, Hongwu, Anatoly Sorokin, Alexander Mazein, Alex Selkov, Evgeni Selkov, Oleg Demin, and Igor Goryanin. 2007. “The Edinburgh Human Metabolic Network Reconstruction and Its Functional Analysis.” *Molecular Systems Biology* 3 (September): 135.
Malm, Magdalena, Rasool Saghaleyni, Magnus Lundqvist, Marco Giudici, Veronique Chotteau, Raymond Field, Paul Varley, et al. n.d. “Evolution from Adherent to Suspension – Systems Biology of HEK293 Cell Line Development.” https://doi.org/10.1101/2020.01.29.924894.
Malm, Magdalena, Rasool Saghaleyni, Magnus Lundqvist, Marco Giudici, Veronique Chotteau, Ray Field, Paul G. Varley, et al. 2020. “Evolution from Adherent to Suspension: Systems Biology of HEK293 Cell Line Development.” *Scientific Reports* 10 (1): 18996.
Mardinoglu, Adil, Rasmus Agren, Caroline Kampf, Anna Asplund, Intawat Nookaew, Peter Jacobson, Andrew J. Walley, et al. 2013. “Integration of Clinical Data with a Genome-Scale Metabolic Model of the Human Adipocyte.” *Molecular Systems Biology* 9 (1): 649.
Mardinoglu, Adil, Rasmus Agren, Caroline Kampf, Anna Asplund, Mathias Uhlen, and Jens Nielsen. 2014. “Genome-Scale Metabolic Modelling of Hepatocytes Reveals Serine Deficiency in Patients with Non-Alcoholic Fatty Liver Disease.” *Nature Communications* 5: 3083.
Martínez, José L., Lifang Liu, Dina Petranovic, and Jens Nielsen. 2012. “Pharmaceutical Protein Production by Yeast: Towards Production of Human Blood Proteins by Microbial Fermentation.” *Current Opinion in Biotechnology* 23 (6): 965–71.
Martorell-Marugán, Jordi, Siham Tabik, Yassir Benhammou, Coral del Val, Igor Zwir, Francisco Herrera, and Pedro Carmona-Sáez. 2019. “Deep Learning in Omics Data Analysis and Precision Medicine.” In *Computational Biology*, edited by Holger Husi. Brisbane (AU): Codon Publications.
Mattanovich, Diethard, Paola Branduardi, Laura Dato, Brigitte Gasser, Michael Sauer, and Danilo Porro. 2012. “Recombinant Protein Production in Yeasts.” *Methods in Molecular Biology* 824: 329–58.
Ma, Yunqi, Chang-Joo Lee, and Jang-Su Park. 2020. “Strategies for Optimizing the Production of Proteins and Peptides with Multiple Disulfide Bonds.” *Antibiotics (Basel, Switzerland)* 9 (9). https://doi.org/10.3390/antibiotics9090541.
McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. “Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation.” *Nucleic Acids Research* 40 (10): 4288–97.
McDermaid, Adam, Brandon Monier, Jing Zhao, Bingqiang Liu, and Qin Ma. 2019. “Interpretation of Differential Gene Expression Results of RNA-Seq Data: Review and Integration.” *Briefings in Bioinformatics* 20 (6): 2044–54.
McKusick, V. A. 1997. “Genomics: Structural and Functional Studies of Genomes.” *Genomics* 45 (2): 244–49.
McLeish, T. C. B. 2005. “Protein Folding in High-Dimensional Spaces: Hypergutters and the Role of Nonnative Interactions.” *Biophysical Journal* 88 (1): 172–83.
Mendoza, Sebastián N., Brett G. Olivier, Douwe Molenaar, and Bas Teusink. 2019. “A Systematic Assessment of Current Genome-Scale Metabolic Reconstruction Tools.” *Genome Biology* 20 (1): 158.
Meng, Chen, Dominic Helm, Martin Frejno, and Bernhard Kuster. 2016. “moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets.” *Journal of Proteome Research* 15 (3): 755–65.
Meng, Jianghui, and Jiafu Wang. 2015. “Role of SNARE Proteins in Tumourigenesis and Their Potential as Targets for Novel Anti-Cancer Therapeutics.” *Biochimica et Biophysica Acta* 1856 (1): 1–12.
Mertens, Bart J. A. 2017. “Transformation, Normalization, and Batch Effect in the Analysis of Mass Spectrometry Data for Omics Studies.” *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*. https://doi.org/10.1007/978-3-319-45809-0_1.
“Metabolic Flux Analysis.” n.d. Accessed June 15, 2021. https://www.springer.com/gp/book/9781493911691.
Meuris, Leander, Francis Santens, Greg Elson, Nele Festjens, Morgane Boone, Anaëlle Dos Santos, Simon Devos, et al. 2014. “GlycoDelete Engineering of Mammalian Cells Simplifies N-Glycosylation of Recombinant Proteins.” *Nature Biotechnology* 32 (5): 485–89.
Micheel, Christine M., Sharly J. Nass, Gilbert S. Omenn, Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes, Board on Health Care Services, Board on Health Sciences Policy, and Institute of Medicine. 2012. *Summary*. National Academies Press (US).
Misra, Biswapriya B., Carl D. Langefeld, Michael Olivier, and Laura A. Cox. 2018. “Integrated Omics: Tools, Advances, and Future Approaches.” *Journal of Molecular Endocrinology*, July. https://doi.org/10.1530/JME-18-0055.

Monte, Federica del, Federica del Monte, and Giulio Agnetti. 2014. “Protein Post-Translational Modifications and Misfolding: New Concepts in Heart Failure.” *PROTEOMICS - Clinical Applications*. https://doi.org/10.1002/prca.201400037.
Mo, Qianxing, Sijian Wang, Venkatraman E. Seshan, Adam B. Olshen, Nikolaus Schultz, Chris Sander, R. Scott Powers, Marc Ladanyi, and Ronglai Shen. 2013. “Pattern Discovery and Cancer Gene Identification in Integrated Cancer Genomic Data.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (11): 4245–50.
Morrow, Thomas, and Linda Hull Felcone. 2004. “Defining the Difference: What Makes Biologics Unique.” *Biotechnology Healthcare* 1 (4): 24–29.
Mossmann, Anelise, Guilherme Luiz Dotto, Dachamir Hotza, Sérgio Luiz Jahn, and Edson Luiz Foletto. 2019. “Preparation of Polyethylene–supported Zero–valent Iron Buoyant Catalyst and Its Performance for Ponceau 4R Decolorization by photo–Fenton Process.” *Journal of Environmental Chemical Engineering* 7 (2): 102963.
Mullard, Asher. 2021. “2020 FDA Drug Approvals.” *Nature Reviews. Drug Discovery* 20 (2): 85–90.
Najafi, Ali, Gholamreza Bidkhori, Joseph H. Bozorgmehr, Ina Koch, and Ali Masoudi-Nejad. 2014. “Genome Scale Modeling in Systems Biology: Algorithms and Resources.” *Current Genomics* 15 (2): 130–59.
Nielsen, Jens. 2017a. “Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine.” *Cell Metabolism* 25 (3): 572–79.
———. 2017b. “Systems Biology of Metabolism.” *Annual Review of Biochemistry* 86 (June): 245–75.
Nielsen, Jens, and Stefan Hohmann. 2017. *Systems Biology*. John Wiley & Sons.
Nilsson, Avlant, Jurgen R. Haanstra, Martin Engqvist, Albert Gerding, Barbara M. Bakker, Ursula Klingmüller, Bas Teusink, and Jens Nielsen. 2020. “Quantitative Analysis of Amino Acid Metabolism in Liver Cancer Links Glutamate Excretion to Nucleotide Synthesis.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (19): 10294–304.
Nilsson, Avlant, and Jens Nielsen. 2016. “Metabolic Trade-Offs in Yeast Are Caused by F1F0-ATP Synthase.” *Scientific Reports* 6 (March): 22264.
Nookaew, Intawat, Michael C. Jewett, Asawin Meechai, Chinae Thammarongtham, Kobkul Laoteng, Supapon Cheevadhanarak, Jens Nielsen, and Sakarindr Bhumiratana. 2008. “The Genome-Scale Metabolic Model iIN800 of Saccharomyces Cerevisiae and Its Validation: A Scaffold to Query Lipid Metabolism.” *BMC Systems Biology* 2 (August): 71.
Nusinow, David P., and Steven P. Gygi. 2020. “A Guide to the Quantitative Proteomic Profiles of the Cancer Cell Line Encyclopedia.” *bioRxiv*. https://doi.org/10.1101/2020.02.03.932384.
O ’brien, Edward J., Joshua A. Lerman, Roger L. Chang, Daniel R. Hyduke, and Bernhard Ø. Palsson. 2013. “Genome-Scale Models of Metabolism and Gene Expression Extend and Refine Growth Phenotype Prediction.” *Molecular Systems Biology* 9 (1). https://doi.org/10.1038/msb.2013.52.
O’Brien, Edward J., Jonathan M. Monk, and Bernhard O. Palsson. 2015. “Using Genome-Scale Models to Predict Biological Capabilities.” *Cell* 161 (5): 971–87.
Obudulu, Ogonna, Niklas Mähler, Tomas Skotare, Joakim Bygdell, Ilka N. Abreu, Maria Ahnlund, Madhavi Latha Gandla, et al. 2018. “A Multi-Omics Approach Reveals Function of Secretory Carrier-Associated Membrane Proteins in Wood Formation Of Populus trees.” *BMC Genomics* 19 (1): 11.
O’Callaghan, Peter M., and David C. James. 2008. “Systems Biotechnology of Mammalian Cell Factories.” *Briefings in Functional Genomics & Proteomics* 7 (2): 95–110.
Oftadeh, Omid, Pierre Salvy, Maria Masid, Maxime Curvat, Ljubisa Miskovic, and Vassily Hatzimanikatis. 2021. “A Genome-Scale Metabolic Model of Saccharomyces Cerevisiae That Integrates Expression Constraints and Reaction Thermodynamics.” *bioRxiv*. https://doi.org/10.1101/2021.02.17.431671.
Opdam, Sjoerd, Anne Richelle, Benjamin Kellman, Shanzhong Li, Daniel C. Zielinski, and Nathan E. Lewis. 2017. “A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models.” *Cell Systems* 4 (3): 318–29.e6.
Orth, Jeffrey D., Ines Thiele, and Bernhard Ø. Palsson. 2010. “What Is Flux Balance Analysis?” *Nature Biotechnology* 28 (3): 245–48.
Ozaki, Toshinori, and Akira Nakagawara. 2011. “Role of p53 in Cell Death and Human Cancers.” *Cancers* 3 (1): 994–1013.
Palsson, Bernhard. 2006. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press.
Palsson, Bernhard Ø. 2011. *Systems Biology: Simulation of Dynamic Network States*. Cambridge University Press.
———. 2015. *Systems Biology: Constraint-Based Reconstruction and Analysis*. Cambridge University Press.
Pandey, Vikash, Noushin Hadadi, and Vassily Hatzimanikatis. 2019. “Enhanced Flux Prediction by

Integrating Relative Expression and Relative Metabolite Abundance into Thermodynamically Consistent Metabolic Models." *PLoS Computational Biology* 15 (5): e1007036.
Parker, Carol E., Maria R. Warren, and Viorel Mocanu. 2011. "Mass Spectrometry for Proteomics." In *Neuroproteomics*, edited by Oscar Alzate. Boca Raton (FL): CRC Press/Taylor & Francis.
Park, Solip, Jae-Seong Yang, Young-Eun Shin, Juyong Park, Sung Key Jang, and Sanguk Kim. 2011. "Protein Localization as a Principal Feature of the Etiology and Comorbidity of Genetic Diseases." *Molecular Systems Biology* 7 (May): 494.
Parola, Cristina, Daniel Neumeier, and Sai T. Reddy. 2018. "Integrating High-Throughput Screening and Sequencing for Monoclonal Antibody Discovery and Engineering." *Immunology* 153 (1): 31–41.
Patro, Rob, Stephen M. Mount, and Carl Kingsford. 2014. "Sailfish Enables Alignment-Free Isoform Quantification from RNA-Seq Reads Using Lightweight Algorithms." *Nature Biotechnology* 32 (5): 462–64.
Pevsner, Jonathan. 2009. *Bioinformatics and Functional Genomics*. Wiley.
Pfau, Thomas, Maria Pires Pacheco, and Thomas Sauter. 2016. "Towards Improved Genome-Scale Metabolic Network Reconstructions: Unification, Transcript Specificity and beyond." *Briefings in Bioinformatics* 17 (6): 1060–69.
Pierre-Jean, Morgane, Jean-François Deleuze, Edith Le Floch, and Florence Mauger. 2020. "Clustering and Variable Selection Evaluation of 13 Unsupervised Methods for Multi-Omics Data Integration." *Briefings in Bioinformatics* 21 (6): 2011–30.
Pisal, Dipak S., Matthew P. Kosloski, and Sathy V. Balu-Iyer. 2010. "Delivery of Therapeutic Proteins." *Journal of Pharmaceutical Sciences* 99 (6): 2557–75.
Pobre, Kristine Faye R., Greg J. Poet, and Linda M. Hendershot. 2019. "The Endoplasmic Reticulum (ER) Chaperone BiP Is a Master Regulator of ER Functions: Getting by with a Little Help from ERdj Friends." *The Journal of Biological Chemistry* 294 (6): 2098–2108.
Pop, Mihai. 2009. "Genome Assembly Reborn: Recent Computational Challenges." *Briefings in Bioinformatics* 10 (4): 354–66.
Presley, J. F., N. B. Cole, T. A. Schroer, K. Hirschberg, K. J. Zaal, and J. Lippincott-Schwartz. 1997. "ER-to-Golgi Transport Visualized in Living Cells." *Nature* 389 (6646): 81–85.
Price, D. L., D. R. Borchelt, and S. S. Sisodia. 1993. "Alzheimer Disease and the Prion Disorders Amyloid Beta-Protein and Prion Protein Amyloidoses." *Proceedings of the National Academy of Sciences of the United States of America* 90 (14): 6381–84.
Quinlan, R. 1993. "4.5: Programs for Machine Learning Morgan Kaufmann Publishers Inc." *San Francisco, USA*.
Quinn, Thomas P., Tamsyn M. Crowley, and Mark F. Richardson. 2018. "Benchmarking Differential Expression Analysis Tools for RNA-Seq: Normalization-Based vs. Log-Ratio Transformation-Based Methods." *BMC Bioinformatics* 19 (1): 274.
Rahimpour, Azam, Behrouz Vaziri, Reza Moazzami, Leila Nematollahi, Farzaneh Barkhordari, Leila Kokabee, Ahmad Adeli, and Fereidoun Mahboudi. 2013. "Engineering the Cellular Protein Secretory Pathway for Enhancement of Recombinant Tissue Plasminogen Activator Expression in Chinese Hamster Ovary Cells: Effects of CERT and XBP1s Genes." *Journal of Microbiology and Biotechnology* 23 (8): 1116–22.
Rajvanshi, Meghna, and Kareenhalli V. Venkatesh. 2013. "Flux Balance Analysis." In *Encyclopedia of Systems Biology*, edited by Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, 749–52. New York, NY: Springer New York.
Ramon, Charlotte, Mattia G. Gollub, and Jörg Stelling. 2018. "Integrating -Omics Data into Genome-Scale Metabolic Network Models: Principles and Challenges." *Essays in Biochemistry* 62 (4): 563–74.
Rath, Oliver, and Frank Kozielski. 2012. "Kinesins and Cancer." *Nature Reviews. Cancer* 12 (8): 527–39.
Raynal, Bertrand, Pascal Lenormand, Bruno Baron, Sylviane Hoos, and Patrick England. 2014. "Quality Assessment and Optimization of Purified Protein Samples: Why and How?" *Microbial Cell Factories* 13 (December): 180.
Reel, Parminder S., Smarti Reel, Ewan Pearson, Emanuele Trucco, and Emily Jefferson. 2021. "Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review." *Biotechnology Advances* 49 (July): 107739.
Robinson, Jonathan L., Amir Feizi, Mathias Uhlén, and Jens Nielsen. 2019. "A Systematic Investigation of the Malignant Functions and Diagnostic Potential of the Cancer Secretome." *Cell Reports* 26 (10): 2622–35.e5.
Robinson, Jonathan L., Pınar Kocabaş, Hao Wang, Pierre-Etienne Cholley, Daniel Cook, Avlant Nilsson, Mihail Anton, et al. 2020. "An Atlas of Human Metabolism." *Science Signaling* 13 (624). https://doi.org/10.1126/scisignal.aaz1482.
Rodosthenous, Theodoulos, Vahid Shahrezaei, and Marina Evangelou. 2020. "Integrating Multi-OMICS Data through Sparse Canonical Correlation Analysis for the Prediction of Complex Traits: A

Comparison Study." *Bioinformatics* 36 (17): 4616–25.
Román, Ramón, Joan Miret, Federica Scalia, Antoni Casablancas, Martí Lecina, and Jordi J. Cairó. 2016. "Enhancing Heterologous Protein Expression and Secretion in HEK293 Cells by Means of Combination of CMV Promoter and IFNα2 Signal Peptide." *Journal of Biotechnology* 239 (December): 57–60.
Rosen, Robert. 1968. "Systems Theory and Biology. Proceedings of the 3rd Systems Symposium, Cleveland, Ohio, Oct. 1966. M. D. Mesarović, Ed. Springer-Verlag, New York, 1968. Xii + 403 Pp., Illus. $16." *Science* 161 (3836): 34–35.
Ross, J. F., and M. Orlowski. 1982. "Growth-Rate-Dependent Adjustment of Ribosome Function in Chemostat-Grown Cells of the Fungus Mucor Racemosus." *Journal of Bacteriology* 149 (2): 650–53.
Russell, Jay H., and Kenneth C. Keiler. 2007. "Peptide Signals Encode Protein Localization." *Journal of Bacteriology* 189 (21): 7581–85.
Saghaleyni, R., M. Malm, J. Zrimec, and R. Razavi. 2020. "Transcriptome Analysis of EPO and GFP HEK293 Cell-Lines Reveal Shifts in Energy and ER Capacity Support Improved Erythropoietin Production in HEK293F Cells." *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.09.16.299966v1.abstract.
Saibil, Helen. 2013. "Chaperone Machines for Protein Folding, Unfolding and Disaggregation." *Nature Reviews. Molecular Cell Biology* 14 (10): 630–42.
Sakaguchi, M. 1997. "Eukaryotic Protein Secretion." *Current Opinion in Biotechnology* 8 (5): 595–601.
Sánchez, Benjamín J., Petri-Jaan Lahtvee, Kate Campbell, Sergo Kasvandik, Rosemary Yu, Iván Domenzain, Aleksej Zelezniak, and Jens Nielsen. 2021. "Benchmarking Accuracy and Precision of Intensity-Based Absolute Quantification of Protein Abundances in Saccharomyces Cerevisiae." *Proteomics* 21 (6): e2000093.
Sánchez, Benjamín J., Cheng Zhang, Avlant Nilsson, Petri-Jaan Lahtvee, Eduard J. Kerkhoven, and Jens Nielsen. 2017. "Improving the Phenotype Predictions of a Yeast Genome-Scale Metabolic Model by Incorporating Enzymatic Constraints." *Molecular Systems Biology* 13 (8): 935.
Schaaij-Visser, Tieneke B. M., Meike de Wit, Siu W. Lam, and Connie R. Jiménez. 2013. "The Cancer Secretome, Current Status and Opportunities in the Lung, Breast and Colorectal Cancer Context." *Biochimica et Biophysica Acta* 1834 (11): 2242–58.
Schena, M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." *Science* 270 (5235): 467–70.
Schinn, Song-Min, Carly Morrison, Wei Wei, Lin Zhang, and Nathan E. Lewis. 2021. "A Genome-Scale Metabolic Network Model and Machine Learning Predict Amino Acid Concentrations in Chinese Hamster Ovary Cell Cultures." *Biotechnology and Bioengineering* 118 (5): 2118–23.
Schulz, Christian, Tjasa Kumelj, Emil Karlsen, and Eivind Almaas. 2021. "Genome-Scale Metabolic Modelling When Changes in Environmental Conditions Affect Biomass Composition." *PLoS Computational Biology* 17 (5): e1008528.
Schwarz, Hubert, Liang Zhang, Niklas Andersson, Bernt Nilsson, and Veronique Chotteau. 2019. "Small-Scale End-to-End mAb Platform with a Continuous and Integrated Design," Integrated Continuous Biomanufacturing IV, . https://dc.engconfintl.org/biomanufact_iv/4/.
Schwarz, Hubert, Ye Zhang, Caijuan Zhan, Magdalena Malm, Raymond Field, Richard Turner, Christopher Sellick, Paul Varley, Johan Rockberg, and Véronique Chotteau. 2020. "Small-Scale Bioreactor Supports High Density HEK293 Cell Perfusion Culture for the Production of Recombinant Erythropoietin." *Journal of Biotechnology* 309 (February): 44–52.
Sergeeva, Daria, Karen Julie la Cour Karottki, Jae Seong Lee, and Helene Faustrup Kildegaard. 2019. "CRISPR Toolbox for Mammalian Cell Engineering." *Cell Culture Engineering*. Wiley. https://doi.org/10.1002/9783527811410.ch8.
Shen, Ronglai, Adam B. Olshen, and Marc Ladanyi. 2009. "Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis." *Bioinformatics* 25 (22): 2906–12.
Silverbush, Dana, Simona Cristea, Gali Yanovich-Arad, Tamar Geiger, Niko Beerenwinkel, and Roded Sharan. 2019. "Simultaneous Integration of Multi-Omics Data Improves the Identification of Cancer Driver Modules." *Cell Systems* 8 (5): 456–66.e5.
Skach, William R. 2007. "The Expanding Role of the ER Translocon in Membrane Protein Folding." *The Journal of Cell Biology* 179 (7): 1333–35.
Sohda, M., Y. Misumi, A. Yano, N. Takami, and Y. Ikehara. 1998. "Phosphorylation of the Vesicle Docking Protein p115 Regulates Its Association with the Golgi Membrane." *The Journal of Biological Chemistry* 273 (9): 5385–88.
Söllner, Thomas H. 2003. "Regulated Exocytosis and SNARE Function (Review)." *Molecular Membrane Biology* 20 (3): 209–20.
Sommer, Christoph, and Daniel W. Gerlich. 2013. "Machine Learning in Cell Biology - Teaching

Computers to Recognize Phenotypes." *Journal of Cell Science* 126 (Pt 24): 5529–39.
Sorzano, C. O. S., J. Vargas, and A. Pascual Montano. 2014. "A Survey of Dimensionality Reduction Techniques." *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1403.2877.
Šoštarić, Nikolina, and Vera van Noort. 2021. "Molecular Dynamics Shows Complex Interplay and Long-Range Effects of Post-Translational Modifications in Yeast Protein Interactions." *PLoS Computational Biology* 17 (5): e1008988.
Spahn, Philipp N., Anders H. Hansen, Henning G. Hansen, Johnny Arnsdorf, Helene F. Kildegaard, and Nathan E. Lewis. 2016. "A Markov Chain Model for N-Linked Protein Glycosylation--towards a Low-Parameter Tool for Model-Driven Glycoengineering." *Metabolic Engineering* 33 (January): 52–66.
Stach, Christopher S., Meghan G. McCann, Conor M. O'Brien, Tung S. Le, Nikunj Somia, Xinning Chen, Kyoungho Lee, et al. 2019. "Model-Driven Engineering of N-Linked Glycosylation in Chinese Hamster Ovary Cells." *ACS Synthetic Biology* 8 (11): 2524–35.
Stanley, Pamela. 2011. "Golgi Glycosylation." *Cold Spring Harbor Perspectives in Biology* 3 (4). https://doi.org/10.1101/cshperspect.a005199.
Stastna, Miroslava, and Jennifer E. Van Eyk. 2012. "Secreted Proteins as a Fundamental Source for Biomarker Discovery." *Proteomics* 12 (4-5): 722–35.
Stein-O'Brien, Genevieve L., Raman Arora, Aedin C. Culhane, Alexander V. Favorov, Lana X. Garmire, Casey S. Greene, Loyal A. Goff, et al. 2018. "Enter the Matrix: Factorization Uncovers Knowledge from Omics." *Trends in Genetics: TIG* 34 (10): 790–805.
Stepanenko, A. A., and V. V. Dmitrenko. 2015. "HEK293 in Cell Biology and Cancer Research: Phenotype, Karyotype, Tumorigenicity, and Stress-Induced Genome-Phenotype Evolution." *Gene*. Elsevier. https://doi.org/10.1016/j.gene.2015.05.065.
Steuer, Ralf. 2007. "Computational Approaches to the Topology, Stability and Dynamics of Metabolic Networks." *Phytochemistry* 68 (16-18): 2139–51.
Stevens, F. J., and Y. Argon. 1999. "Protein Folding in the ER." *Seminars in Cell & Developmental Biology* 10 (5): 443–54.
Štor, Jerneja, David E. Ruckerbauer, Diana Széliová, Jürgen Zanghellini, and Nicole Borth. 2021. "Towards Rational Glyco-Engineering in CHO: From Data to Predictive Models." *Current Opinion in Biotechnology* 71 (May): 9–17.
Subramanian, Indhupriya, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. 2020. "Multi-Omics Data Integration, Interpretation, and Its Application." *Bioinformatics and Biology Insights* 14 (January): 1177932219899051.
Sun, Yan V., and Yi-Juan Hu. 2016. "Integrative Analysis of Multi-Omics Data for Discovery and Functional Studies of Complex Human Diseases." *Advances in Genetics* 93 (January): 147–90.
Supplitt, Stanislaw, Pawel Karpinski, Maria Sasiadek, and Izabela Laczmanska. 2021. "Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine." *International Journal of Molecular Sciences* 22 (3). https://doi.org/10.3390/ijms22031422.
Tambuyzer, Erik, Benjamin Vandendriessche, Christopher P. Austin, Philip J. Brooks, Kristina Larsson, Katherine I. Miller Needleman, James Valentine, et al. 2020. "Therapies for Rare Diseases: Therapeutic Modalities, Progress and Challenges Ahead." *Nature Reviews. Drug Discovery* 19 (2): 93–111.
Tarca, Adi L., Vincent J. Carey, Xue-Wen Chen, Roberto Romero, and Sorin Drăghici. 2007. "Machine Learning and Its Applications to Biology." *PLoS Computational Biology* 3 (6): e116.
Tavassoly, Iman, Joseph Goldfarb, and Ravi Iyengar. 2018. "Systems Biology Primer: The Basic Methods and Approaches." *Essays in Biochemistry* 62 (4): 487–500.
Taverna, Domenico, and Marco Gaspari. 2021. "A Critical Comparison of Three MS-Based Approaches for Quantitative Proteomics Analysis." *Journal of Mass Spectrometry: JMS* 56 (1): e4669.
Tegel, Hanna, Melanie Dannemeyer, Sara Kanje, Åsa Sivertsson, Anna Berling, Anne-Sophie Svensson, Andreas Hober, et al. 2020. "High Throughput Generation of a Resource of the Human Secretome in Mammalian Cells." *New Biotechnology* 58 (September): 45–54.
Thak, Eun Jung, Su Jin Yoo, Hye Yun Moon, and Hyun Ah Kang. 2020. "Yeast Synthetic Biology for Designed Cell Factories Producing Secretory Recombinant Proteins." *FEMS Yeast Research* 20 (2). https://doi.org/10.1093/femsyr/foaa009.
Tharwat, Alaa, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. 2017. "Linear Discriminant Analysis: A Detailed Tutorial." *AI Communications. The European Journal on Artificial Intelligence* 30 (2): 169–90.
The UniProt Consortium. 2017. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45 (D1): D158–69.
Thiele, Ines, Neema Jamshidi, Ronan M. T. Fleming, and Bernhard Ø. Palsson. 2009. "Genome-Scale Reconstruction of Escherichia Coli's Transcriptional and Translational Machinery: A Knowledge Base,

Its Mathematical Formulation, and Its Functional Characterization." *PLoS Computational Biology* 5 (3): e1000312.
Thiele, Ines, and Bernhard Ø. Palsson. 2010. "A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction." *Nature Protocols* 5 (1): 93–121.
Thiele, Ines, Neil Swainston, Ronan M. T. Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K. Aurich, Hulda Haraldsdottir, et al. 2013. "A Community-Driven Global Reconstruction of Human Metabolism." *Nature Biotechnology* 31 (5): 419–25.
Thul, Peter J., Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, et al. 2017. "A Subcellular Map of the Human Proteome." *Science* 356 (6340). https://doi.org/10.1126/science.aal3321.
Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society* 58 (1): 267–88.
Toby, Timothy K., Luca Fornelli, and Neil L. Kelleher. 2016. "Progress in Top-Down Proteomics and the Analysis of Proteoforms." *Annual Review of Analytical Chemistry* 9 (1): 499–519.
Töpfer, Nadine, Sabrina Kleessen, and Zoran Nikoloski. 2015. "Integration of Metabolomics Data into Metabolic Networks." *Frontiers in Plant Science* 6 (February): 49.
Töpfer, Nadine, Samuel M. D. Seaver, and Asaph Aharoni. 2018. "Integration of Plant Metabolomics Data with Metabolic Networks: Progresses and Challenges." In *Plant Metabolomics: Methods and Protocols*, edited by Carla António, 297–310. New York, NY: Springer New York.
Tsai, Billy, Yihong Ye, and Tom A. Rapoport. 2002. "Retro-Translocation of Proteins from the Endoplasmic Reticulum into the Cytosol." *Nature Reviews. Molecular Cell Biology* 3 (4): 246–55.
Uhlén, Mathias, Linn Fagerberg, Bjö M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419–1260419.
Uhlén, Mathias, Max J. Karlsson, Andreas Hober, Anne-Sophie Svensson, Julia Scheffel, David Kotol, Wen Zhong, et al. 2019. "The Human Secretome." *Science Signaling* 12 (609). https://doi.org/10.1126/scisignal.aaz0274.
Vcelar, Sabine, Vaibhav Jadhav, Michael Melcher, Norbert Auer, Astrid Hrdina, Rebecca Sagmeister, Kelley Heffner, et al. 2018. "Karyotype Variation of CHO Host Cell Lines over Time in Culture Characterized by Chromosome Counting and Chromosome Painting." *Biotechnology and Bioengineering* 115 (1): 165–73.
Verissimo, Fatima, and Rainer Pepperkok. 2013. "Imaging ER-to-Golgi Transport: Towards a Systems View." *Journal of Cell Science* 126 (Pt 22): 5091–5100.
Vieira Gomes, Antonio Milton, Talita Souza Carmo, Lucas Silva Carvalho, Frederico Mendonça Bahia, and Nádia Skorupa Parachin. 2018. "Comparison of Yeasts as Hosts for Recombinant Protein Production." *Microorganisms* 6 (2). https://doi.org/10.3390/microorganisms6020038.
Visser, E. K., C. G. van Reenen, J. T. N. van der Werf, M. B. H. Schilder, J. H. Knaap, A. Barneveld, and H. J. Blokhuis. 2002. "Heart Rate and Heart Rate Variability during a Novel Object Test and a Handling Test in Young Horses." *Physiology & Behavior* 76 (2): 289–96.
Voit, Eberhard O. 2013. *A First Course in Systems Biology*. Garland Science.
Volkova, Svetlana, Marta R. A. Matos, Matthias Mattanovich, and Igor Marín de Mas. 2020. "Metabolic Modelling as a Framework for Metabolomics Data Integration and Analysis." *Metabolites* 10 (8). https://doi.org/10.3390/metabo10080303.
Walker, Lary C., and Harry LeVine 3rd. 2012. "Corruption and Spread of Pathogenic Proteins in Neurodegenerative Diseases." *The Journal of Biological Chemistry* 287 (40): 33109–15.
Walker, Stephen J., and Mark O. Lively. 2013. "Chapter 778 - Signal Peptidase (Eukaryote)." In *Handbook of Proteolytic Enzymes (Third Edition)*, edited by Neil D. Rawlings and Guy Salvesen, 3512–17. Academic Press.
Wang, Bo, Vivek Kumar, Andrew Olson, and Doreen Ware. 2019. "Reviving the Transcriptome Studies: An Insight Into the Emergence of Single-Molecule Transcriptome Sequencing." *Frontiers in Genetics* 10 (April): 384.
Wang, Hao, Simonas Marcišauskas, Benjamín J. Sánchez, Iván Domenzain, Daniel Hermansson, Rasmus Agren, Jens Nielsen, and Eduard J. Kerkhoven. 2018. "RAVEN 2.0: A Versatile Toolbox for Metabolic Network Reconstruction and a Case Study on Streptomyces Coelicolor." *PLoS Computational Biology* 14 (10): e1006541.
Wang, Jianbo, Zhenqing Ye, Tim H-M Huang, Huidong Shi, and Victor Jin. 2015. "A Survey of Computational Methods in Transcriptome-Wide Alternative Splicing Analysis." *Biomolecular Concepts* 6 (1): 59–66.
Wang, Shiyu, and Randal J. Kaufman. 2012. "The Impact of the Unfolded Protein Response on Human Disease." *The Journal of Cell Biology* 197 (7): 857–67.
Wang, Shuyu, Zhi-Yang Tsun, Rachel L. Wolfson, Kuang Shen, Gregory A. Wyant, Molly E. Plovanich,

Elizabeth D. Yuan, et al. 2015. “Metabolism. Lysosomal Amino Acid Transporter SLC38A9 Signals Arginine Sufficiency to mTORC1.” *Science* 347 (6218): 188–94.
Wang, Willam T., Breeya A. Tailor, David S. Cohen, and Xudong Huang. 2019. “Alzheimer’s Pathogenesis, Metal-Mediated Redox Stress, and Potential Nanotheranostics.” *EC Pharmacology and Toxicology* 7 (7): 547–58.
Wang, Yong, and Nicholas E. Navin. 2015. “Advances and Applications of Single-Cell Sequencing Technologies.” *Molecular Cell* 58 (4): 598–609.
Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews. Genetics* 10 (1): 57–63.
Weis, Roland, Franz Hartner, and Anton Glieder. 2006. “Recombinant Protein Production in Yeast.” In *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*, 1620–25. Berlin, Heidelberg: Springer Berlin Heidelberg.
Welsh, John B., Lisa M. Sapinoso, Suzanne G. Kern, David A. Brown, Tao Liu, Asne R. Bauskin, Robyn L. Ward, et al. 2003. “Large-Scale Delineation of Secreted Protein Biomarkers Overexpressed in Cancer Tissue and Serum.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (6): 3410–15.
Wilk-Blaszczak, Malgosia. n.d. “Cell Physiology.” Accessed May 4, 2021. https://uta.pressbooks.pub/cellphysiology/chapter/posttranslational-modifications-of-proteins/.
Wilks, D. S. 2011. “Chapter 13 - Canonical Correlation Analysis (CCA).” In *International Geophysics*, edited by Daniel S. Wilks, 100:563–82. Academic Press.
Wilson, Christopher M., Kaiqiao Li, Xiaoqing Yu, Pei-Fen Kuan, and Xuefeng Wang. 2019. “Multiple-Kernel Learning for Genomic Data Mining and Prediction.” *BMC Bioinformatics* 20 (1): 426.
Wong, Chi-Huey. 2005. “Protein Glycosylation: New Challenges and Opportunities.” *The Journal of Organic Chemistry* 70 (11): 4219–25.
Worton, R. G., C. C. Ho, and C. Duff. 1977. “Chromosome Stability in CHO Cells.” *Somatic Cell Genetics* 3 (1): 27–45.
Wurm, Florian. 2013. “CHO Quasispecies—Implications for Manufacturing Processes.” *Processes* 1 (3): 296–311.
Wu, Thomas D., and Colin K. Watanabe. 2005. “GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences.” *Bioinformatics* 21 (9): 1859–75.
Wyant, Gregory A., Monther Abu-Remaileh, Rachel L. Wolfson, Walter W. Chen, Elizaveta Freinkman, Laura V. Danai, Matthew G. Vander Heiden, and David M. Sabatini. 2017. “mTORC1 Activator SLC38A9 Is Required to Efflux Essential Amino Acids from Lysosomes and Use Protein as a Nutrient.” *Cell* 171 (3): 642–54.e12.
Xia, Jianye, Benjamin Sánchez, Yu Chen, Kate Campbell, Sergo Kasvandik, and Jens Nielsen. 2021. “Proteome Allocations Change Linearly with Specific Growth Rate of Saccharomyces Cerevisiae under Glucose-Limitation.” Research Square. https://doi.org/10.21203/rs.3.rs-464207/v1.
Xing, Zizhuo, Brian M. Kenty, Zheng Jian Li, and Steven S. Lee. 2009. “Scale-up Analysis for a CHO Cell Culture Process in Large-Scale Bioreactors.” *Biotechnology and Bioengineering* 103 (4): 733–46.
Xu, Chengchao, and Davis T. W. Ng. 2015. “Glycosylation-Directed Quality Control of Protein Folding.” *Nature Reviews. Molecular Cell Biology* 16 (12): 742–52.
Yang, Lei, Yingli Lv, Tao Li, Yongchun Zuo, and Wei Jiang. 2014. “Human Proteins Characterization with Subcellular Localizations.” *Journal of Theoretical Biology* 358 (October): 61–73.
Yan, Kang K., Hongyu Zhao, and Herbert Pang. 2017. “A Comparison of Graph- and Kernel-Based –omics Data Integration Algorithms for Classifying Complex Traits.” *BMC Bioinformatics* 18 (1). https://doi.org/10.1186/s12859-017-1982-4.
Yizhak, Keren, Tomer Benyamini, Wolfram Liebermeister, Eytan Ruppin, and Tomer Shlomi. 2010. “Integrating Quantitative Proteomics and Metabolomics with a Genome-Scale Metabolic Network Model.” *Bioinformatics* 26 (12): i255–60.
Yoshida, Hiderou. 2007. “ER Stress and Diseases.” *The FEBS Journal* 274 (3): 630–58.
Zampieri, Guido, Supreeta Vijayakumar, Elisabeth Yaneske, and Claudio Angione. 2019. “Machine and Deep Learning Meet Genome-Scale Metabolic Modeling.” *PLoS Computational Biology* 15 (7): e1007084.
Zampieri, Mattia, Karthik Sekar, Nicola Zamboni, and Uwe Sauer. 2017. “Frontiers of High-Throughput Metabolomics.” *Current Opinion in Chemical Biology* 36 (February): 15–23.
Zhai, Ya-Qi, Ning-Li Chai, Hui-Kai Li, Zhong-Sheng Lu, Xiu-Xue Feng, Wen-Gang Zhang, Sheng-Zhen Liu, and En-Qiang Linghu. 2020. “Endoscopic Submucosal Excavation and Endoscopic Full-Thickness Resection for Gastric Schwannoma: Five-Year Experience from a Large Tertiary Center in China.” *Surgical Endoscopy* 34 (11): 4943–49.
Zhang, Li, Chenkai Lv, Yaqiong Jin, Ganqi Cheng, Yibao Fu, Dongsheng Yuan, Yiran Tao, Yongli Guo, Xin Ni, and Tieliu Shi. 2018. “Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic

Subtypes in High-Risk Neuroblastoma." *Frontiers in Genetics* 9 (October): 477.
Zhang, Shihua, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W. Laird, and Xianghong Jasmine Zhou. 2012. "Discovery of Multi-Dimensional Modules by Integrative Analysis of Cancer Genomic Data." *Nucleic Acids Research* 40 (19): 9379–91.
Zhang, Weijing, Weiling He, Yongjie Shi, Haifeng Gu, Min Li, Zhimin Liu, Yanling Feng, Nianzhen Zheng, Chuanmiao Xie, and Yanna Zhang. 2016. "High Expression of KIF20A Is Associated with Poor Overall Survival and Tumor Progression in Early-Stage Cervical Squamous Cell Carcinoma." *PloS One* 11 (12): e0167449.